Les statistiques chez Wikipedia

Joseph Saint Pierre

rencontre ingénieurs statisticiens 11 juillet 2013

Le 12 novembre 2012 j'ai reçu un message d'une enseignante de l'université de Toulouse le Mirail me demandant de l'aide pour le traitement statistiques une table de contingence, il s'agissait d'une table 2 par 2 et l'une des 4 valeurs théoriques était inférieure à valeur 5, le test χ^2 de Pearson est considéré comme non approprié et certains logiciels émettent une mise en garde et proposent d'utiliser d'autres tests, c'est notamment le cas pour **SPSS**. Un des tests souvent proposés par les logiciels pour ce type de situation est le test exact de Fisher. Comme ce test est facile à calculer et apparemment facile à comprendre j'ai conseillé l'utilisation de celui-ci et plutôt que d'écrire une explication sur le test j'ai cherché une bonne présentation sur Internet et j'en ai trouvé une excellente, la page de l'encyclopédie libre wikipedia en langue anglaise, j'avais lancé la recherche en français mais ce que j'avais trouvé dans cette langue ne me convenait pas. Il y avait aussi une page en allemand, en italien, en russe, en suédois, en japonais, en basque, j'ai étudié allemand et j'arrive à comprendre les mathématiques en italien, j'ai trouvé étrange l'absence de page wikipedia en français sur ce test très simple. Après quelques jours de réflexions et quelques discussions je me suis dit qu'il ne serait pas très difficile de créer cette page en commençant par une traduction sommaire de la page en anglais, en allant prendre quelques idées sur les pages en allemand et en italien tout en faisant appel à mes connaissances sur les logiciels de statistiques et à mes souvenirs de cours de statistiques.

Entre le 26 et le 30 novembre 2012 en 5 étapes j'ai composé l'essentiel ce qui constitue la page wikipedia sur ce test. Pour information et afin d'en garer une trace voici le texte repris de la page wikipedia le 11 juin 2013.

Le Test exact de Fisher est un test statistique utilisé pour l'analyse des tables de contingence. Ce test est utilisé en général avec des faibles effectifs mais il est valide pour toutes les tailles d'échantillon. Il doit son nom à son inventeur Ronald Fisher. C'est un test qualifié d'exact car les probabilités peuvent être calculées exactement plutôt qu'en s'appuyant sur une approximation qui ne devient correcte qu'asymptotiquement comme pour le test du χ^2 utilisé dans les tables de contingence.

Les calculs à la main ne sont raisonnables que pour les tables 2 par 2 mais le principe du test peut s'étendre au cas général et certains logiciels de statistique permettent le calcul pour le cas général.

Soit un échantillon d'adolescents on sépare l'échantillon entre filles et garçons et entre ceux qui suivent un régime et ceux qui n'en suivent pas et nous supposons que la proportion de filles qui suivent un régime est supérieure à celle des garçons, et nous voulons tester si la différence de proportions observées est significative. Voici les données

	Garçons	Filles	total ligne
régime	1	9	10
non régime	11	3	14
total colonne	12	12	24

Ces données ne sont pas adaptée pour une analyse par un test du χ^2, parce que les valeurs attendues (théoriques) dans la table sont inférieures à 10, et dans une table de contingence 2 par 2, le nombre de degrés de liberté est toujours égal à 1.

La question que l'on se pose à propos de ces données est : sachant que 10 de ces 24 adolescents pratiquent un régime et que 12 sont des filles, quelle est la probabilité que ces 10 qui pratiquent un régime soient répartis de manière équilibrée entre les filles et les garçons ? Si on choisit 10 adolescents au hasard, quelle est la probabilité que 9 d'entre eux soient parmi les 12 filles et seulement 1 parmi les 12 garçons ?

Avant de passer au test de Fisher nous introduisons quelques notations. On représente les cellules par les lettres a, b, c et d et on note n le total général. La table se présente ainsi :

	Garçons	Filles	total ligne
régime	a	b	a+b
non régime	c	d	d
total colonne	a+c	b+d	a+b+c+d (=n)

Fisher a montré que la probabilité d'obtenir un tel ensemble de valeurs était donnée par la loi hypergéometrique :

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

où $\binom{n}{k}$ est le coefficient binomial et le symbole ! indique la factorielle.

Dans l'exemple la probabilité d'obtenir le tableau croisé observé, avec les totaux marginaux donnés, est donc :

$$p = \frac{10!14!12!12!}{1!9!11!3!24!} = \frac{40}{29716} = 0,001346...$$

De manière à calculer si des données observées sont significativement éloignées de l'indépendance, c'est-à-dire la probabilité d'observer des données aussi ou plus éloignées

que celles observées si l'hypothèse nulle (indépendance) est satisfaite, il faut calculer les valeurs de p pour ces tables et les ajouter. Cela donne un test unilatéral ; pour un test bilatéral on doit considérer les tables qui sont extrêmes mais dans l'autre direction. Malheureusement classer les tables pour savoir si elles sont ou non aussi extrêmes est problématique. Une approche utilisée dans **R** avec la fonction "fisher.test" calcule la valeur p en sommant les probabilités de toutes les tables ayant une probabilité inférieure ou égale à celle de la table observée.

Voici la commande permettant d'entrer le tableau croisé de l'exemple présenté plus haut comme une matrice de **R** et de calculer la valeur de la probabilité.

$fisher.test(matrix(c(1, 9, 11, 3), nrow = 2))$

Et voici la valeur de la probabilité obtenue :

$p - value = 0.002759$

Cette valeur se calcule aussi assez facilement sans logiciel puissant avec le principe exposé. Il y a en tout quatre tables aussi éloignées ou plus éloignées de l'indépendance que la table observée. Il n'y a en tout que onze tables possibles sachant qu'il y a seulement dix adolescents qui suivent un régime, le nombre de garçons suivant un régime peut varier de 0 jusqu'à 10. Une fois que le nombre de garçons suivant un régime est connu tout le reste de la table est connu, ce qui correspond intuitivement au degré de liberté valant 1. Les quatre tables extrêmes correspondent aux valeurs 0,1,9 et 10 pour le nombre de garçons suivant un régime. La probabilité des deux tables plus éloignées de l'indépendance est très faible elle se calcule comme précédemment :

$p = \frac{10!14!12!12!}{0!10!12!2!24!} = \frac{1}{29716} = 0,0000336519...$

En ajoutant les probabilités des quatre tables on trouve

$p = \frac{82}{29716} = 0,0027594561...$

Le test permet de rejeter l'indépendance entre le sexe et le fait de faire un régime.

Comme cela a été noté plus haut la plupart des logiciels de statistiques modernes calculent le niveau de signification du test exact de Fisher, même si l'approximation du test du χ^2 est acceptable. Les calculs faits par les logiciels de statistiques sont différents des règles expliquées plus haut en raison des difficultés numériques dues aux très grandes valeurs des factorielles. Une meilleure approche s'appuie sur une fonction gamma ou une fonction log-gamma, mais les méthodes de calcul précis pour les probabilités hypergéométriques ou binomiales font encore partie de la recherche.

Attention, il n'y a que le contenu proprement dit de la page et ne figurent pas ici les liens créés plus ou moins automatiquement par le système de l'encyclopédie ni les liens créés par les utilisateurs, contributeurs. J'ai passé plus de temps à regarder les liens, à tenter d'en corriger certains qu'à écrire la page. La page wikipedia en français sur le test du χ^2 mentionnait le test de exact de Fisher tout en mettant un lien sur une page qui traite d'un autre test de Fisher sur la comparaison des variances, ce test suit bien une loi

F de Fisher mais ne correspond pas au test usuel en régression, analyse de variance. La mise en place du lien que j'estime correct, sur la page que j'ai créée m'a obligé a entrer dans une polémique visible sur wikipedia, malgré le nom le test exact de Fisher est assez éloigné des tests de Fisher utilisant la loi F.

J'ai regardé beaucoup de pages en français et en anglais sur les statistiques et je me suis dit que je pouvais me livrer à un travail de traduction adaptation à partir des pages en anglais bien meilleures. Cela m'a très fortement rappelé mes deux ans de recherche après ma thèse entre 1984 et 1986, j'étais impliqué dans un projet de recherche européen dont le but était de comparer l'utilisation des statistiques en sciences sociales entre la France et le monde anglophone. Après avoir écrit la page j'ai eu un échange en anglais avec 3 de mes anciens collègues, notamment le directeur, australien, du laboratoire où j'étais et un collègue qui m'a appris énormément de choses sur les logiciels, les ordinateurs etc. Je vous livre l'essentiel du message qui a initié la discussion :

Recently a lecturer in social sciences sent me a message about a 2 by 2 contingency table with small expected values. She thought chi-square was not suitable and ask me for another test. I proposed to use Fisher's exact test and I tried ton find a good explanation of this test on Internet. I could not find any good page in french and the english wikipedia on this test seems to me quite good. There was a wikipedia page in german, in italian, in japanese, in swedish, in russian, in basque. I thought I could start writing a french wikipedia page on this test, I started writing the page on november 26th with a rough and partial translation of the english page :

`http://en.wikipedia.org/wiki/Fisher%27s_exact_test`
`http://fr.wikipedia.org/wiki/Test_exact_de_Fisher`

According to wikipedia principles I hope this page will be improved by others statisticians.

Translating a page on statistics from english to french like this one remains me strongly the research project in which I was involved beteween 1984 and 1986 in Lancaster. Not only because of translation but mainly the contingency tables stuff. SPAD package I started using in Lancaster has some computations based on Fisher's exact test.

But the most important thing is the enduring differences between french and "british" styles and ways of presenting mathematical things. I did not translate the part on "lady testing tea" which is a very good introduction in the english page. I guess that in a "true" french version (not a translation) hypergeometric distribution would be presented first and Fisher's excat test would be presented as an application of this distribution.

In 1985 I became fellow of Royal Statistical Society and as I receive the « significance »journal (statistics making sense). In december 2012 volume 9 issue 6 I read a brilliant article « R.A. Fisher, a lady, and a nice cup of tea »or « Tea for three - of infusions and inferences and milk in first »page 30-33. I really enjoyed reading this paper.

It can be found there :

`http://www.significancemagazine.org/details/magazine/3818261/Tea-for-three-Of-infu`
`html`

I thought I have heard about the author and searching the Internet I found that Stephen Senn is a very well known statistician.

I am still working as statistician in Toulouse university

Yours faithfully

Joseph
`http://cict.fr/~stpierre/`

Ce message initial m'a valu des réponses, des discussions qui m'ont incité à préparer cet exposé.

Comme je l'ai écrit en anglais je n'ai pas traduit la partie historique très intéressante sur « la dame goûtant le thé ». Fisher a eu l'idée du test en buvant du thé et cette anecdote est très célèbre et racontée par Fisher dans son livre très important de 1935 « The Design of Experiments »

Un livre sur l'histoire des statistiques au vingtième siècle paru en 2001 a pour titre : « The Lady Tasting Tea : How Statistics Revolutionized Science in the Twentieth Century »

`http://en.wikipedia.org/wiki/The_Lady_Tasting_Tea`

Les réflexions sur le rôle des statistiques, leur histoire sont bien plus importantes dans les pays de langue anglaise qu'en France. Une différence que j'ai pointée concerne la place du calcul des probabilités. Les pages en français de l'encyclopédie wikipedia sur les lois de probabilités m'ont paru très bonnes en général et ce fut le cas pour la page sur la loi hyper-géométrique qui est assez ancienne, a de nombreux contributeurs. De manière générale les pages sur les mathématiques me semblent correctes.

La page en français sur le test exact de Fisher n'a pratiquement pas été modifiée entre le 30 novembre 2012 et le 17 juin 2013, alors que la page en anglais a été modifiée, sa création remonte à 2004 et a eu 127 contributeurs. La page en allemand date de 2006 et compte plus de 20 contributeurs différents, la page italienne date aussi de 2006 mais je n'ai pas évalué le nombre de contributeurs. Pour certaines langues il est possible d'obtenir des comptage de consultations voici un petit tableau pour donner un ordre d'idée sur le nombre de consultations pour trois langues, anglais, italien et français sur la période novembre 2012 à mai 2013.

Mois	Anglais	Italien	Français
Nov 2012	25020	912	84
Déc 2012	19933	785	294
Jan 2013	22801	1135	438
Fév 2013	24548	1095	526
Mar 2013	28474	1028	805
Avr 2013	30135	955	1256
Mai 2013	27665	1089	1385

La comparaison de ces chiffres n'est pas simple car il faudrait relativiser par rapport aux nombres de visites sur l'ensemble des pages de l'encyclopédie par langues. Cela donnerait des tables de contingence difficile à analyser finement.

Il est par contre simple de comparer le nombre de fois où la page en français sur le test exact de Fisher a été accédée avec le nombre de fois où des documents figurant sur ma page web l'ont été. Par comparaison le document le plus télé-chargé en mai 2013 sur ma page est la documentation sur le logiciel libre PSPP, il y a eu 201 accès. Mais il s'agit d'un document long qui a demandé 3 mois de stage pour être écrit, très sérieux et très riche, lu et relu.

Lors d'une précédente intervention dans le cadre des rencontres entre ingénieurs statisticiens toulousains le 13 mars 2012 j'avais présenté le logiciel libre PSPP. Il se trouve que la documentation sur ce logiciel libre est en grande partie liée à mes réflexions sur l'encyclopédie libre wikipedia. En effet une part importante des accès au document sur ma page web provient des pages wikipedia sur PSPP en anglais, en français, en espagnol et en japonais sur lesquelles figurent des liens vers ma page web.

Il y a plusieurs statuts possibles par rapport à l'encyclopédie libre, et jusqu'au début 2007 j'ignorais celui correspondant à trouver dans l'encyclopédie des liens sur ma page. De plus ce qui était écrit dans l'encyclopédie en français sur le logiciel PSPP me semblait très éloigné de ma vision du logiciel et de son intérêt.

J'avais envisagé de reprendre fortement cette page mais cela soulevait et soulève encore des questions non résolues. La popularité de l'encyclopédie libre wikipedia tend à rendre tous ses articles fortement consultés, lus, assimilés, voire pris comme références. L'existence de cette encyclopédie et surtout sa richesse pourrait induire le raisonnement suivant « Pourquoi former des spécialistes puis qu'on peut trouver des expertises sur wikipedia ? »Dans le prolongement de « Pourquoi former au calcul puisqu'il y a des machines pour cela ? »Après l'ordinateur, les logiciels, les systèmes, les encyclopédies en ligne seraient une étape supplémentaire pour justifier une absence d'apprentissage.

Un autre point important lié à la nature libre de l'encyclopédie concerne le travail des contributeurs celui-ci est peu visible et les lecteurs ne sont pas identifiés. Répondre à un enseignant, ou un étudiant en conseillant l'usage d'une méthode et donnant quelques explications est une activité aisément identifiée.

La réponse que j'ai donnée en conseillant l'utilisation du test exact de Fisher n'est pas nécessairement la meilleure réponse possible face à une table de contingence, il me semble que souvent je conseillerai plutôt l'utilisation du test du rapport de vraisemblance qui est beaucoup plus général, mais sans doute plus difficile à expliquer rapidement. Je n'ai pas trouvé sur wikipedia en français une bonne explication des tests et le cas du test du rapport de vraisemblance incompréhensible. Le travail du statisticien consiste donc souvent encore à trier les informations.

En regardant certaines statistiques sur les fréquentations, les nombres d'articles par langue ou par pays où se trouvent les contributeurs on observe quelques faits surprenants

comme l'importance du néerlandais. Les sept langues ayant le plus d'articles, plus d'un million en mai 2013, sont dans l'ordre l'anglais suivi de l'allemand, du néerlandais, du français, de l'italien, de l'espagnol et du russe. Les pages ne sont pas nécessairement écrites par quelqu'un dont c'est la première langue ou la langue majoritaire du pays. Il y a des pages en latin et en Yiddish... L'encyclopédie Wikipedia est organisée en langues et non en pays, wikipedia en français n'est pas une encyclopédie pour les français mais pour les francophones. Il y a bien un site `http://wikipedia.fr` c'est celui de l'association Wikimedia France, celle ci n'a pas vraiment d'autorité sur le contenu de l'encyclopédie en français pas plus que Wikimedia Canada ou Wikimedia Suisse. En avril 2013 le président de Wikimedia France, Rémi Mathis a été convoqué par la Direction Centrale du Renseignement Intérieur pour lui demander de supprimer la page en français sur la Station hertzienne militaire de Pierre-sur-Haute. La suppression a eu lieu mais elle a été temporaire la page a été traduite en plus de 20 langues, la page en français a été écrite à nouveau depuis un autre pays que la France.

Les considérations sur les positions particulières de la France sur Internet m'ont amené à regarder les positions sur l'accord commercial anti-contrefaçon plus connu sous son acronyme anglais ACTA (Anti-Counterfeiting Trade Agreement). Cet accord a déclenché une violente opposition des partisans des logiciels libres. En 2012 il y a eu vote au parlement européen sur cet accord Sur 754 députés, seul 39 ont voté en faveur de cet accord et parmi ceux là il y avait 21 députés français. Sachant qu'il y a 74 députés français au parlement européen il est possible de reconstituer un tableau de contingence 2 par 2 pour mesurer la liaison entre vote pour ACTA et appartenance à la France, les autres modalités pour le vote abstention et opposition son regroupées. Il y a eu plus de 150 abstentions mais je n'ai pas retrouvé les abstention par pays, j'ai trouvé sur plusieurs sites la liste de ceux qui ont voté pour.

	ACTA	Abs+Contre	total ligne
France	21	53	74
Reste Europe	18	662	680
total colonne	39	715	754

On peut ici calculer le test exact de Fisher, il s'agit bien d'un exemple où une valeur théorique est inférieure à 5 et où l'usage du test de χ^2 est déconseillé

$fisher.test(matrix(c(21, 53, 18, 662), nrow = 2))$

$p - value = 6.547e - 13$

Le test est très hautement significatif. Le vote en faveur de ACTA est fortement lié aux français. On peut toutefois remarquer que l'appartenance politique joue un rôle essentiel. L'immense majorité des députés ayant voté en faveur d'ACTA font partie du même groupe appelé Parti Populaire Européen, qui comprend des partis comme comme la CDU Allemande, le Parti Populaire Espagnol ou l'UMP Française mais pas le Parti Conservateur Britannique. Il y a 33 députés de ce groupe ayant voté pour ACTA ; voici la table correspondant :

	ACTA	Abs+Contre	total ligne
PPE	33	236	269
Autres	6	479	485
total colonne	39	715	754

On peut aussi calculer le test exact de Fisher
$fisher.test(matrix(c(33, 236, 6, 479), nrow = 2))$

$p - value = 1.455e - 10$

Ici aussi le test exact de Fisher est très hautement significatif, il semble évident que le vote en faveur d'ACTA est fortement lié à l'appartenance au groupe PPE. Un regard attentif sur les données montre que les 21 députés français qui ont voté pour l'ACTA fort parti du groupe du Parti Populaire Européen alors qu'il y a 6 députés européens non français qui ont voté pour l'ACTA et qui n'appartiennent pas au Parti Populaire Européen et donc seulement 12 qui y appartiennent. Cette différence suggère qu'il y a une interaction d'ordre 3 entre le vote pour ACTA, l'appartenance à la France et l'appartenance au Parti Populaire. Une des faiblesses du test exact de Fisher est de ne pas être, au moins pour le moment, aux tables de contingence à 3 entrées ou plus. Sur ce petit jeu de données j'ai pu mesurer l'effet significatif de cet interaction avec une régression logistique faite dans SAS avec la proc GENMOD :

```
data acta ;
input pays \$ groupe \$ vota dept ;
cards ;
Eur DC 12 239
Eur ap 6 441
France DC 21 30
France ap 0 44
;
proc genmod;
class pays groupe ;
model vota/dept=  groupe * pays
                 /dist =bin link=logit waldci type3 obstat;
run;
```

```
                    LR Statistics For Type 3 Analysis

                                      Chi-
            Source              DF    Square     Pr > ChiSq

            pays*groupe          3    111.65        <.0001
```

Je n'ai jamais vraiment exploré comment produire ce type de calcul avec R même si je pense que cela doit être possible.

L'interaction entre le pays et le groupe est significative, le vote pour ACTA est spécifique de la France, spécifique du Parti Populaire Européen et il y a un effet supplémentaire montrant une association particulièrement forte entre la composante française du Parti Populaire Européen et le vote en faveur d'ACTA.

Le modèle loglinéaire mais aussi les arbres de décision peuvent mettre en évidence ce genre de modèles avec des interactions.

Il est possible de percevoir cette interaction avec le calcul du test exact de Fisher pour quatre tables de contingence deux par deux.
une pour les députés français, en croisant le groupe avec le vote
une pour les députés non français, en croisant le groupe avec le vote
une pour les député du groupe PPE ,en croisant le pays avec le vote
une pour les députés hors groupe PPE,en croisant le pays avec le vote

Voici ce que donne le calcul pour ces quatres tables

$fisher.test(matrix(c(21, 9, 0, 44), nrow = 2))$

$p - value = 9.446e - 12$

$fisher.test(matrix(c(12, 227, 6, 435), nrow = 2))$

$p - value = 0.009857$

$fisher.test(matrix(c(21, 9, 12, 227), nrow = 2))$

$p - value = 3.543e - 16$

$fisher.test(matrix(c(12, 227, 6, 435), nrow = 2))$

$p - value = 0.009857$

Le test du rapport de vraisemblance sur ces quatre tables de contingence montre la même chose, heureusement. La liaison entre le vote pour ACTA et l'appartenance partisane est extrêmement forte chez les français et significative mais nettement moins chez les autres pays et cela malgré des effectifs nettement plus faible pour les français. La différence la plus forte dans le vote ACTA est entre les députés français du goupe PPE et les députés non français du même groupe. Il n'y par contre aucune différence significative entre les députés français et ceux des autres pays pour les autres groupes que le PPE. Ces deux faits sont l'indice d'une interaction entre les trois variables pays, groupe et vote.

La régression logistique, le modèle loglinéaire permettent d'étudier les interactions dans le cadre plus général des modéles linéaires généralisé. Vous pouvez truver ma thèse sur ma page web, où figure une présentation un peu théorique de ces modèles :

http://cict.fr/~stpierre/thesen.pdf

J'ai écrit un exposé pour des doctorants sur ce que l'on nomme abusivement le paradoxe de Simpson, sur un fichier très similaire avec aussi trois variables à deux modalités. Il est disponible ici :

http://cict.fr/~stpierre/expose-16-01-98/expose.html

et au format pdf ici :

http://cict.fr/~stpierre/simpson.pdf

Le test exact de Fisher permet de voir un lien pas trop compliqué entre une loi de probabilité et une méthode statistique pour identifier des liaisons entre variables catégorielles mais il ne permet pas facilement de déboucher vers une réflexion sur la modélisation des tables de contingences, c'est pour cela que la rédaction de la page était très facile. Des pages isolées sur des sujets ponctuels ne peuvent pas constituer, à mon avis, un bon cours sur les statistiques.

Comme je l'ai mentionné dans le message en anglais que j'ai reproduit, le test exact de Fisher est utilisé dans SPAD et cela de manière exploratoire pour détecter des liaisons entre modalités de variables qualitatives, en utilisant un principe que j'ai utilisé au dessus pour étudier les liens entre votes, pays et partis, c'est à dire regrouper les modalités des variables pour n'avoir que deux modalités, en gardant une modalité d'intérêt et regroupant les autres, il y a 27 pays 8 partis et 3 votes et je n'ai regardé que des tableaux deux par deux sur lesquels le test exact de Fisher est facilement calculable. Il y a presque 30 ans (1984) que j'ai pu utiliser ces méthodes et elles me semblent avoir une pertinence mais si elles ont des limites mentionnées précédemment ou beaucoup plus lourdes d'un point de vue Bayesien.

Outre l'usage exploratoire du test exact de Fisher dans SPAD on peut trouver d'autres tentaives plus récentes en liaison avec les graphes comme ici :

http://ugrad.stat.ubc.ca/R/library/GeneTS/html/00Index.html

Ce test léger à mettre en œuvre semble peut avoir un usage pour les traitements de données massives.

Hormis un, les exemples choisis précédemment pour cet exposé donne des tests exacts de Fisher significatifs, or l'exemple historique sur la dame goûtant le thé ne l'était pas. À partir des données sur le vote ACTA j'ai pu noter une toute petite chose, le pays qui a voté le plus pour cet accord après la France est l'Allemagne avec 8 votes. Les 8 députés qui ont voté en faveur de l'accord appartiennent tous au groupe parti populaire du parlement européen mais ces 8 députés n'appartient pas tous au même parti ni au même land, 3 députés ayant voté pour ACTA appartiennent au parti Union chrétienne-sociale en Bavière (Christlich-Soziale Union in Bayern) connu sous le sigle CSU et 5 députés appartiennent au parti Union chrétienne-démocrate d'Allemagne (Christlich Demokratische Union Deutschlands) dont le sigle est CDU. Il y a en tout 8 députés de la CSU et 34

députés de la CDU. La proportion de députés bavarois ayant voté pour l'accord ACTA est forte, plus que celle des députés allemands non bavarois.

$fisher.test(matrix(c(3, 5, 5, 29), nrow = 2))$

$p - value = 0.1625$

En conclusion, il est difficile de porter un jugement simple sur la qualité des statistiques sur wikipedia en juillet 2013. L'encyclopédie évolue rapidement et son histoire est assez récente, l'évolution future dépendra des contributions et donc des contributeurs. Une de mes contributions récentes pour le café politique a été intitulée :

« Internet sera ce qu'en feront les Internautes, son histoire n'est pas écrite »elle se trouve ici :

`http://lecafepolitique.free.fr/spip.php?article288`

Cela pourrait se transposer à l'encyclopédie libre wikipedia mais aussi aux logiciels libres. J'ai abordé le sujet de l'encyclopédie libre dans un texte de réflexion sur les logiciels libres :

`http://cict.fr/~stpierre/logiciels-libres.html`

De manière générale les encyclopédies comme les dictionnaires peuvent souffrir d'un découpage en articles, certes avec des renvois plus ou moins nombreux. Il n'est pas toujours facile de constituer à partir de cela une vision globale, une cartographie des méthodes. Et à partir de cela il n'est pas facile de reconstituer l'ensemble d'une démarche statistique dans un traitement de données.

Wikipedia s'est globalement améliorée depuis son démarrage, malgré certains défauts mentionnées dans ce document il est possible d'être raisonablement optimiste sur l'évolution future.

Toutes les réflexions écrites dans ce document correspondent à la situatin observée en juillet 2013 et seront vraisemblablement modifiées.