A propos d'un test d'indépendance

Le problème évoqué page 148 conduit au tableau suivant :

	A	B	Total
S	4956	1835	6791
NS	1862	668	2530
Total	6818	2503	9321

Les fréquences observées sont par conséquent données par :

	A	B	Total
S	0,5317	0,1969	0,7286
NS	0,1998	0,0717	0,2714
Total	0,7315	0,2685	1

1- Imaginons que A et S soient indépendants et les probabilités p et q de A et de S soient connus (cela peut arriver si on connaît bien les répartitions des abonnements A et B pour chacun des 10000 acheteurs ainsi que la répartition N, NS tout en ignorant les valeurs des quatre cases du tableau) alors :

$$\delta^2 = \left(\frac{(f_{AS} - pq)^2}{pq} + \frac{(f_{BS} - (1-p)q)^2}{(1-p)q} + \frac{(f_{ANS} - p(1-q))^2}{p(1-q)} + \frac{(f_{BNS} - (1-p)(1-q))^2}{(1-p)(1-q)} \right),$$

où f_{AS}, f_{BS}, f_{ANS} et f_{BNS} sont les fréquences de résultats dans A∩S, B∩S, A∩NS et B∩NS, est petite et diminue avec la taille de l'échantillon ; de façon plus précise $n\ \delta^2$ suit asymptotiquement une loi du χ^2 à 3 degrés de liberté. On serait ramené à un test d'adéquation à une loi sur un ensemble à 4 éléments représenté par le tableau suivant :

	A	B	Total
S	$p\ q$	$(1-p)\ q$	q
NS	$p\ (1-q)$	$(1-p)\ (1-q)$	$1-q$
Total	p	$1-p$	1

2- Considérons maintenant n simulations d'une loi P quelconque définie par le tableau suivant, dont aucune case n'a la valeur 0 :

	A	B	Total
S	p_{AS}	p_{BS}	p_S
NS	p_{ANS}	p_{BNS}	p_{NS}
Total	p_A	p_B	1

et considérons la « distance » :

$$d^2 = \frac{(f_{AS} - f_A f_S)^2}{f_A f_S} + \frac{(f_{BS} - f_B f_S)^2}{f_B f_S} + \frac{(f_{ANS} - f_A f_{NS})^2}{f_A f_{NS}} + \frac{(f_{BNS} - f_B f_{NS})^2}{f_B f_{NS}}.$$

où f_{AS}, f_{BS}, f_{ANS}, f_{BNS}, f_A, f_S, f_B, f_{NS} sont les fréquences de résultats dans A∩S, B∩S, A∩NS, B∩NS, A, S, B et NS, cette expression ayant l'avantage de ne faire intervenir aucun paramètre théorique de P.

> On peut simplifier l'expression de d^2 en remarquant que :
> $(f_{AS} - f_A f_S)^2 = (f_{AS}(f_{AS} + f_{ANS} + f_{BS} + f_{BNS}) - (f_{AS} + f_{ANS})(f_{AS} + f_{BS}))^2$
> $= (f_{AS} f_{BNS} - f_{ANS} f_{BS})^2$, résultat que l'on obtient aussi avec $(f_{AS} - f_A f_S)^2$, $(f_{BS} - f_B f_S)^2$, $(f_{ANS} - f_A f_{NS})^2$ et $(f_{BNS} - f_B f_{NS})^2$.
>
> Dans ces conditions :
> $$d^2 = (f_{AS} f_{BNS} - f_{ANS} f_{BS})^2 \left(\frac{1}{f_A f_S} + \frac{1}{f_B f_S} + \frac{1}{f_A f_{NS}} + \frac{1}{f_B f_{NS}} \right)$$
> $$= \frac{(f_{AS} f_{BNS} - f_{ANS} f_{BS})^2}{f_A f_S f_B f_{NS}}$$

Que penser de d^2 ?
d^2 est une variable aléatoire discrète qui peut prendre la valeur + ∞. On peut conjecturer que, lorsque n augmente, d^2 tend vers $l = \frac{(p_{AS} p_{BNS} - p_{ANS} p_{BS})^2}{p_A p_S p_B p_{NS}}$, cette limite l étant nulle si et seulement si A et S sont des événements indépendants (simulation correspondante à l'adresse : http://perso.wanadoo.fr/jpq/proba/test-independance/abni.htm).

3- De plus, il apparaît (grâce à des simulations réalisées à l'aide de http://perso.wanadoo.fr/jpq/proba/test-independance/ par exemple), et cela se démontre, que, sous l'hypothèse d'indépendance, pour n grand, la répartition de $n\,d^2$ ne dépend ni de n ni de p et q ; la loi limite s'appelle loi du χ^2 à 1 degré de liberté. Pour tout autre modèle ne vérifiant pas « A et S sont indépendants », $n\,d^2$ tend vers l'infini. La mesure de $n\,d^2$ est donc propre à distinguer le modèle avec indépendance de A et S des autres.

Dans le problème qui nous occupe, une simulation de 9321 expériences conduit à un dernier décile pour d^2 de l'ordre de 0,00030.

Le d^2 observé est ici $\frac{(4956 \times 668 - 1862 \times 1835)^2}{6818 \times 2503 \times 6791 \times 2530}$ = 0,000038438 et donc on ne peut rejeter l'hypothèse d'indépendance avec un risque d'erreur de 10%.

N.B. : on peut aussi remarquer que 9321 d^2 est de l'ordre de 2,71 (valeur donnée par les tables du khi2 et comparer alors 9321 d^2_{obs} à ce seuil.

4- L'hypothèse d'indépendance étant maintenant acceptée, toute loi P (loi de probabilité sur un ensemble à 4 éléments) modélisant la situation en jeu est complètement déterminée par P(A) et P(B) ; à partir des données dont on dispose, on pourra prendre

P(A) = 0,7315 et P(S) = 0,7286 et choisir pour modéliser cette situation la loi définie dans le tableau ci-dessous.

	A	B	Total
S	0,5329	0,1957	0,7286
NS	0,1986	0,0728	0,2714
Total	0,7315	0,2685	1