est, pour n grand, distribuée selon une loi qui ne dépend que de k. Cette loi s'appelle loi du khi deux à $k-1$ degrés de liberté et on trouve dans des tables numériques la liste des 9e déciles de ces lois (voir tableau ci-dessous). On peut ainsi généraliser la méthode ci-dessus et définir un critère de compatibilité d'une série de données avec une loi quelconque sur un ensemble fini.

$k-1$	1	2	3	4	5	6	10	20	30
$\alpha = 0{,}1$	2,71	4,61	6,25	7,78	9,24	10,64	15,99	28,41	40,26
$\alpha = 0{,}05$	3,84	5,99	7,81	9,49	11,07	12,59	18,31	31,41	43,77

Pour $k = 2$, on part d'une loi binomiale et les calculs peuvent se retrouver autrement (voir sur le cédérom, « Compléments aux documents d'accompagnement »).
L'objectif ici n'est pas que les élèves fassent eux-mêmes la simulation, mais qu'ils soient capables de définir une règle de décision et d'exploiter les résultats de simulations.

Test d'indépendance

Dans le paragraphe « Probabilités conditionnelles et indépendance », on s'est posé la question de l'indépendance des variables abonnement et statut pour lesquelles on dispose du tableau suivant, donnant les résultats de ce couple de variables sur $N = 9321$ individus (voir tableau (1) ci-dessous).

	A	B
S	4956	1835
NS	1862	668

Tableau (1)

La traduction dans le champ de la statistique de cette question est : peut-on trouver un modèle compatible avec les données, défini par deux nombres p et r tels que les probabilités des 4 événements en jeu soient celles qui sont données dans le tableau ci-dessous :

	A	B
S	pr	$p(1-r)$
NS	$(1-p)r$	$(1-p)(1-r)$

Si tel est le cas, la quantité d^2 suivante doit être *petite* :

$$d^2 = \left(\frac{4956}{9321} - pr\right)^2 + \left(\frac{1835}{9321} - p(1-r)\right)^2 + \left(\frac{1862}{9321} - (1-p)r\right)^2 + \left(\frac{668}{9321} - (1-p)(1-r)\right)^2.$$

Mais on ne connaît ni p ni r. Un objectif en statistique est ici de trouver une fonction des données d'un tableau tel le (1) dont la répartition se stabilise, lorsque le nombre de données devient grand, vers une répartition qui ne dépend ni de p ni de r ; c'est le cas pour la fonction définie ci-dessous, les données du tableau (2) étant remplacées par des lettres (tableau (2')) :

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

Tableau (2)

	A	B	Totaux
S	a	b	n
NS	c	d	N-n
Totaux	m	N-m	N

Tableau(2')

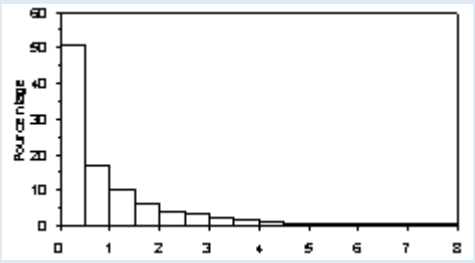
$$z = N\left(\frac{(a' - \hat{p}\hat{r})}{\hat{p}\hat{r}} + \frac{(b' - \hat{p}(1-\hat{r}))^2}{\hat{p}(1-\hat{r})} + \frac{(c' - (1-\hat{p})\hat{r})^2}{(1-\hat{p})\hat{r}} + \frac{(d' - (1-\hat{p})(1-\hat{r}))^2}{(1-\hat{p})(1-\hat{r})}\right)$$

où : $\hat{p} = \frac{n}{N}, \hat{r} = \frac{m}{N}, a' = \frac{a}{N}, \ldots, d' = \frac{d}{N}$

(on remarque que s'il existe un modèle compatible avec les données, défini par p et r, où les événements A et S sont indépendants, alors $\hat{p} = \frac{n}{N}$ et $\hat{r} = \frac{m}{N}$ seront voisins de p et r).

On peut écrire z plus simplement sous la forme suivante : $z = \dfrac{N(ad-bc)^2}{nm(N-n)(N-m)}$.

On démontre en théorie des probabilités que la répartition asymptotique (*i.e.* pour N tendant vers $+\infty$) des valeurs de z ne dépend ni de p et r ; en pratique, pourvu que N soit suffisamment grand et p et r pas trop voisins de 0 ou 1, on approchera la loi de probabilité de z par la loi limite, à savoir la loi du χ^2 à 1 degré de liberté. Pour avoir l'allure de cette répartition simulons des tableaux avec $N = 1000$, $p = r = 1/2$ et calculons pour chaque tableau la valeur de z.



Histogramme de 2000 valeurs de z

On voit que 90 % des valeurs de la série simulée sont inférieures à 2,7 ; on peut convenir de la règle de décision suivante :
– si la valeur observée de z est $\leqslant 2{,}7$, alors les variables en jeu seront dites indépendantes ;
– si la valeur observée de z est $> 2{,}7$, alors les variables en jeu seront dites non indépendantes.
On associera à cette conclusion le risque $\alpha = 0{,}1$ correspondant au fait suivant : en utilisant cette règle de décision sur les données simulées, on se serait trompé dans 10 % des cas.
Si la valeur observée de z est inférieure ou égale à 2,7, on pourra choisir le modèle défini par :

$$P(A \text{ et } S) = \hat{p}\hat{r},\ P(B \text{ et } S) = \hat{p}(1-\hat{r}),\ P(A \text{ et } NS) = (1-\hat{p})\hat{r},\ P(B \text{ et } NS) = (1-\hat{p})(1-\hat{r}).$$

Pour le tableau (1) la valeur observée de z est 0,36 : les variables en jeu sont dites indépendantes au risque 0,1, ou au niveau de confiance 1-0,1 = 0,9. On pourra prendre le modèle suivant :

$$P(A \text{ et } S) = \hat{p}\hat{r} = 0{,}53,\ P(B \text{ et } S) = \hat{p}(1-\hat{r}) = 0{,}20,\ P(A \text{ et } NS) = (1-\hat{p})\hat{r} = 0{,}20,\ P(B \text{ et } NS) = (1-\hat{p})(1-\hat{r}) = 0{,}07\ .$$

La démarche suivie est donc de tester l'hypothèse qu'il existe un modèle où A et S sont indépendants et qui est compatible avec les données. Si cette hypothèse est acceptable, on construit alors une loi P qui vérifie $P(\text{A et S}) = P(\text{A})P(\text{S})$.

Cette conclusion suppose que les données manquantes ne masquent pas un phénomène spécifique ; ainsi, si les 679 cas exclus au départ sont tous des non salariés qui prennent l'abonnement B, alors la valeur observée de z est 225 et la conclusion change !

Statistique et TICE

Le développement rapide de l'usage de la statistique est lié à celui de l'informatique. Pour une sensibilisation à la statistique, dans le cadre d'un enseignement de mathématiques, il convient cependant de cerner en quoi les outils logiciels sont indispensables. Il ne s'agit pas d'initier les élèves à un logiciel spécialisé de statistique, ni même de les entraîner à utiliser systématiquement les possibilités de logiciels comme les tableurs ou les logiciels de géométrie (il serait utile que les enseignants acquièrent une bonne maîtrise de tels outils). On pourra se limiter à quelques possibilités indispensables à une mise en œuvre efficace des programmes de seconde, première et terminale.
On distinguera notamment les usages suivants :
– calculs simples, tels ceux de moyenne, d'écart type qui peuvent être faits sur calculatrice par un ordinateur quasi-instantanément même sur des séries de grandes tailles ;