Ajustement de données expérimentales à une loi équirépartie(1)

Michel Henry

Introduction

Les programmes de terminale S et de terminale ES en vigueur à la rentrée 2002 comportent un paragraphe, identique dans les deux filières, intitulé *Simulation* et formulé comme suit :

« Étude d'un exemple traitant de l'adéquation de données expérimentales à une loi équirépartie ».

Le commentaire est le suivant :

« L'élève devra être capable de poser le problème de l'adéquation à une loi équirépartie et de se reporter à des résultats de simulation qu'on lui fournit. Le vocabulaire des tests (test d'hypothèses, hypothèse nulle, risque de première espèce) est hors programme ».

C'est bien entendu aux tests du Khi-deux (appelés ainsi car ils font intervenir la loi d'une variable standard notée χ^2) qu'il est fait allusion.

L'introduction de cette notion avant le baccalauréat est une nouveauté, dont on ne peut contester l'intérêt : tester la validité d'un modèle est une démarche essentielle dans toute activité scientifique que beaucoup de nos élèves (y compris ceux qui se dirigeront vers les sciences humaines) auront à pratiquer. Mais elle est délicate à mettre en œuvre, car elle suppose de la part des enseignants une vision claire autant de la théorie sous-jacente que de ses enjeux didactiques. Or beaucoup de collègues sont mal à l'aise avec la statistique inférentielle, n'ayant pas eu dans leur cursus la formation *adéquate*.

Objet de la statistique inférentielle ou statistique mathématique

Statistique inférentielle = étude d'une population statistique et prise de décisions à partir de l'observation d'un échantillon.

Notamment :

Estimations de paramètres (proportions, moyennes, écarts-types), exemple des sondages.

Contrôles d'hypothèses (comparaisons de moyennes, ajustements à des lois probabilistes), exemple des tests d'adéquation.

Lorsque les échantillons observés sont aléatoires, la théorie probabiliste permet de contrôler les risques de prendre de mauvaises décisions.

(1) Pour cet atelier, je me suis largement inspiré de l'article *Tests d'adéquation à une loi de probabilité du Khi-deux*, publié avec Louis-Marie Bonneval pour la brochure APMEP n° 156 *Statistique au Lycée*, éditée en octobre 2005 par la Commission Inter-IREM Statistique et probabilités.

Principe d'un test d'hypothèse

Par principe, un test statistique consiste à considérer une hypothèse (au sens de conjecture) dite nulle, notée H_0, et à regarder si l'échantillon prélevé réalise un événement E qui, si l'hypothèse H_0 était vraie, serait de probabilité $P_{H_0}(E)$ relativement petite, inférieure à un seuil donné α.

Dans le cas où l'échantillon observé réalise E, on refuse de considérer que cet événement est seulement dû aux fluctuations d'échantillonnage et on préfère conclure que H_0 n'est pas acceptable (on dit qu'on rejette cette hypothèse, au profit d'une hypothèse alternative H_1).

Ce faisant, on prend un risque de rejeter H_0 alors que seules les fluctuations d'échantillonnage sont responsables de la réalisation de l'événement E, bien qu'il soit peu probable.

Test d'adéquation à une loi de probabilité

Les valeurs observées d'un caractère C défini sur une population statistique P sont souvent réparties entre k modalités M_i qui constituent une partition du domaine de variations de ce caractère. C'est directement le cas quand C est un caractère qualitatif. Il en est de même quand C est quantitatif discret ou continu, l'étude de la répartition de ses valeurs possibles se limite alors à leur distribution entre un nombre fini de classes M_i dont l'explicitation constitue le modèle général dans lequel se situera l'étude.

À partir d'un échantillon de taille n prélevé dans la population P, le principe d'un test d'adéquation consiste à comparer la distribution des fréquences observées $(f_i)_{1\leq i\leq k}$, notée (f), des différentes modalités M_i du caractère C, avec une loi de probabilité *théorique* $(p_i)_{1\leq i\leq k}$, notée (p) par la suite. Cette loi de probabilité constitue un *sous-modèle probabiliste* censé représenter les variations du caractère C dans la population entre les différentes modalités M_i.

Cadre probabiliste du test

On se place dans l'hypothèse où p_i serait la probabilité qu'un élément pris au hasard dans la population P soit de modalité M_i.

Tenant compte des fluctuations d'échantillonnage, les fréquences observées f_i sont considérées comme les valeurs prises sur l'échantillon observé par une famille de variables aléatoires F_i, notée (F) par la suite, définies sur l'ensemble des échantillons aléatoires de taille n que l'on peut prélever dans la population P. Le test apprécie la proximité des deux familles (f) et (p), et, le cas échéant, permet de conclure à *l'adéquation de la distribution des fréquences observées* (f) *à la loi théorique donnée* (p).

Pour pouvoir appliquer des résultats probabilistes puissants, on considère donc que l'échantillonnage est aléatoire (observation de n éléments pris au hasard dans la population, sans que leur prélèvement modifie sensiblement les probabilités des différentes modalités). On suppose aussi que les éléments prélevés le sont indépendamment les uns des autres.

D'un échantillon à l'autre, les fréquences observées fluctuent, reflétant les aléas des prélèvements. Elles respectent cependant la contrainte $\sum_{i=1}^{k} f_i = 1$. On dit que la distribution de fréquences (f) a $k-1$ *degrés de liberté*.

On suppose donc que p_i est effectivement la probabilité de la modalité M_i. On sait d'expérience que quand la taille n de l'échantillon augmente, les fréquences observées f_i *tendent à se stabiliser* vers les probabilités p_i.

Les lois des F_i peuvent d'ailleurs être précisées. En effet, le nombre $N_i = nF_i$ d'éléments de l'échantillon qui sont de modalité M_i suit une loi binomiale $B(n, p_i)$, d'espérance np_i et de variance $np_i(1-p_i)$. L'espérance mathématique de $F_i = \dfrac{N_i}{n}$ est donc p_i et son écart-type est $\sqrt{\dfrac{p_i(1-p_i)}{n}}$.

Le théorème de Bernoulli (loi faible des grands nombres) formalise ce phénomène de stabilisation sous la forme :

$$\forall \varepsilon > 0,\ P(|F_i - p_i| > \varepsilon) \to 0 \text{ quand } n \to \infty.$$

Dans un test d'adéquation, on teste l'hypothèse :

– H_0 : « les probabilités des modalités M_i sont valablement modélisées par la loi de probabilité (p) ».

contre l'hypothèse

– H_1 : « l'écart observé entre la distribution de fréquences (f) des modalités M_i dans l'échantillon et la loi (p) n'est pas dû au hasard du prélèvement de cet échantillon ».

Ainsi, sous l'hypothèse H_0, les p_i sont les valeurs moyennes autour desquelles les fréquences F_i fluctuent.

Mise en œuvre d'un test d'adéquation

Une loi modèle est donc donnée par une famille finie de probabilités (p) où l'on suppose que p_i est la probabilité que le caractère d'un élément pris au hasard dans la population soit de modalité M_i. Les M_i peuvent être les différentes qualités possibles du caractère C ou les différentes classes entre lesquelles on a regroupé les valeurs possibles de C, quand ce caractère est quantitatif.

Un test d'adéquation consiste alors à regarder si, une distance d dans l'espace $\mathbf{R}^k$ étant choisie, la distance $d((f),(p))$ de la loi (p) à la distribution de fréquences (f) observée dans un échantillon de taille n, est inférieure ou supérieure à une certaine *valeur critique* d_c.

Cela revient à définir la valeur critique d_c par la condition que, sous l'hypothèse H_0, la variable aléatoire $D = d((F),(p))$, fonction des variables F_i, ne devrait dépasser d_c qu'avec une probabilité inférieure ou égale à un *seuil de signification* α. On voit que le calcul de cette probabilité de contrôle nécessite la connaissance, au moins approximativement, de la loi de D sous cette hypothèse H_0. C'est le cas

(asymptotiquement[2]) de la distance des tests du Khi-deux (d'autres tests utilisent d'autres distances, comme celui de Kolmogorov-Smirnov).

Distance du Khi-deux

Comme distance d dans $\mathbf{R}^k$, il peut sembler naturel de considérer la distance euclidienne dont le carré est $\sum_{i=1}^{k}(f_i - p_i)^2$, ce serait effectivement une mesure de la dispersion des F_i autour de leurs moyennes p_i. Or, pour les tests du Khi-deux, il s'avère plus intéressant de considérer la quantité

$$d_n^2((f),(p)) = n\sum_{i=1}^{k}\frac{(f_i - p_i)^2}{p_i} = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i} = \left(\sum_{i=1}^{k}\frac{n_i^2}{np_i}\right) - n$$

où $n_i = nf_i$ est l'effectif de la modalité M_i dans l'échantillon.

Cette fonction d peut jouer le rôle d'une distance puisqu'elle est positive et d'autant plus petite que les f_i sont proches des p_i. Mais le fait de pondérer les termes de cette somme par les $\frac{1}{p_i}$ a pour effet de la *normer*, en ce sens que pour n assez grand, la loi de la variable D_n^2 induite ne dépend pratiquement plus de n et des p_i, mais seulement de k. C'est d'ailleurs exactement le cas de sa moyenne, car pour tout n on a :

$$E(D_n^2) = nE\left(\sum_{i=1}^{k}\frac{(F_i - p_i)^2}{p_i}\right) = n\sum_{i=1}^{k}\frac{\mathrm{Var}(F_i)}{p_i} = n\sum_{i=1}^{k}\frac{p_i(1-p_i)}{np_i} = k-1.$$

Une autre raison est que l'on connaît cette loi, c'est asymptotiquement une loi du Khi-deux. Cette propriété fait l'objet du Théorème du Khi-deux, dû à Karl Pearson (1900), conséquence d'un grand théorème probabiliste, le théorème-limite central.

Théorème du Khi-deux

Si, dans un échantillon aléatoire de taille n, F_i désigne la fréquence de la modalité M_i, et si p_i est la probabilité qu'un élément pris au hasard dans la population soit de modalité M_i, la suite des variables aléatoires $D_n^2 = n\sum_{i=1}^{k}\frac{(F_i - p_i)^2}{p_i}$ *converge en loi vers la loi du Khi-deux à $k-1$ degrés de liberté, loi d'une variable notée* χ_{k-1}^2.

Pratiquement, cela signifie que pour *n assez grand*, et pour tout $Q > 0$, les probabilités $P(D_n^2 > Q)$ sont assez proches des $P(\chi_{k-1}^2 > Q)$. Une table du Khi-deux

(2) Quand la taille n de l'échantillon tend vers l'infini, la loi de D^2 converge vers une loi de χ^2 (i.e. La suite des fonctions de répartition des variables aléatoires D^2 converge simplement vers la fonction de répartition de ce χ^2).

(ou la fonction KHIDEUX.INVERSE d'Excel) donne le quantile Q pour un seuil α fixé, vérifiant $\mathrm{P}\left(\chi^2_{k-1} > \mathrm{Q}\right) = \alpha$.

On admet en général que *n assez grand* veut dire : pour tout *i* de 1 à *k*, $np_i(1-p_i) \geq 5$ (on trouve aussi $n > 30$ et $np_i \geq 5$), conditions qui, si elles ne sont pas vérifiées, impliquent de regrouper des modalités pour pouvoir appliquer ce théorème.
La loi de χ^2_ν est bien connue des probabilistes, elle est présentée dans les manuels universitaires de statistique. On a $\mathrm{E}\left(\chi^2_\nu\right) = \nu$ et $\mathrm{Var}\left(\chi^2_\nu\right) = 2\nu$. D'ailleurs, d'après le calcul ci-dessus, pour tout n, on a : $\mathrm{E}\left(\mathrm{D}^2_n\right) = k-1$.

Pratique d'un test du Khi-deux

Dans la pratique, pour faire un test du Khi-deux, une loi modèle conjecturée (*p*) étant donnée, il faut :

1) Vérifier que pour tous les *i* de 1 à *k*, on a $np_i(1-p_i) \geq 5$ (éventuellement regrouper des modalités).

2) Un seuil α étant donné, trouver dans la table du χ^2_{k-1} à $k-1$ degrés de liberté la valeur $\chi^2_c(\alpha)$ telle que $\mathrm{P}\left(\chi^2_{k-1} > \chi^2_c(\alpha)\right) = \alpha$.

3) Expliciter la distribution (*f*) des fréquences présentées par l'échantillon observé.

4) Calculer la *distance* du Khi-deux :

$$\mathrm{d}^2_n = n\sum_{i=1}^{k}\frac{(f_i-p_i)^2}{p_i} = \sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i} = \left(\sum_{i=1}^{k}\frac{n_i^2}{np_i}\right) - n.$$

5) Si $\mathrm{d}^2_n > \chi^2_c(\alpha)$, conclure que la loi (*p*) n'est pas un modèle suffisamment adéquat pour représenter la distribution de fréquences (*f*), en prenant un risque (de *première espèce*) de se tromper $\mathrm{P}_{(p)}\left(\mathrm{D}^2_n > \chi^2_c(\alpha)\right) = \alpha$ inférieur à α.

Adéquation à une loi équirépartie

On veut savoir si on peut représenter la distribution de fréquences (*f*) entre les *k* modalités M_i par un modèle d'équiprobabilité : dans l'hypothèse H_0, les p_i sont donc supposées égales à $\frac{1}{k}$.

L'expression de la *distance* du Khi-deux devient :

$$\mathrm{d}^2_n = nk\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2 = \frac{k}{n}\sum_{i=1}^{k}\left(n_i - \frac{n}{k}\right)^2 = \frac{k}{n}\left(\sum_{i=1}^{k}n_i^2\right) - n.$$

Une remarque peut alors simplifier la pratique de ce test et donner lieu un énoncé simple, accessible en terminale.

Avec $\alpha = 0{,}05$, valeur courante, on peut voir dans la table du Khi-deux que pour tout k, la valeur critique $\chi_c^2(\alpha)$ est comprise entre k et $2k$. Cette simplification, relativement grossière, a l'avantage d'éviter d'utiliser cette table du Khi-deux. Le test consiste alors à comparer la *distance* d_n^2 à $2k$, ce qui minimise le risque de se tromper, car $P_{H_0}\left(D_n^2 > 2k\right) < P_{H_0}\left(D_n^2 > \chi_c^2(\alpha)\right)$.

En simplifiant par k, cela donne la décision pratique :

Si $nk\sum_{i=1}^{k}\left(f_i - \frac{1}{k}\right)^2 > 2$, alors $d_n^2 > \chi_c^2(\alpha)$ et on rejette l'hypothèse d'équirépartition du caractère entre ses k modalités, avec un risque de se tromper $P_{H_0}\left(D_n^2 > 2k\right)$ inférieur à 0,05.

Avec les effectifs observés dans les classes M_i, $n_i = nf_i$, cette condition de rejet de l'équirépartition s'écrit aussi :

$$\sum_{i=1}^{k}\left(n_i - \frac{n}{k}\right)^2 = \left(\sum_{i=1}^{k} n_i^2\right) - \frac{n^2}{k} > 2n.$$

Exemples

1 – Un dé pipé ?

Tester si un dé est régulier.

$$p_i = 1/6,\ np_i(1-p_i) \geq 5 \Rightarrow n \geq 36,\ \chi_c^2(0{,}05) = 11\ ;\ \chi_c^2(0{,}01) = 15.$$

2 – Tester si une table de chiffres au hasard est convenable.

$$p_i = 1/10,\ np_i(1-p_i) \geq 5 \Rightarrow n \geq 56,$$
$$\chi_c^2(0{,}05) = 17\ ;\ \chi_c^2(0{,}01) = 21{,}7\ ;\ \chi_c^2(0{,}1) = 14{,}7.$$

3 – Tester un générateur aléatoire de chiffres. Simuler un générateur pipé, avec différents seuils et tailles d'échantillons.

Éléments de bibliographie

Commission Inter-IREM Statistique et probabilités : *Statistique au Lycée*, brochure APMEP nº 156, octobre 2005.

Commission Inter-IREM Statistique et probabilités : *Autour de la modélisation en probabilités*, Presses universitaires de Franche-Comté, 2001.

Dress, F. : *Probabilités, Statistique, rappels de cours, questions de réflexion, exercices d'entraînement*, Dunod, Paris, 1997.

Dutarte P. : *Pour une éducation à l'inférence statistique au lycée*, Repères-IREM nº 60, juillet 2005.

Dutarte P. : *L'induction statistique au lycée illustrée par le tableur*, Didier, 2005.

Dutarte P., Piednoir J.-L. : *Enseigner la statistique au lycée : des enjeux aux méthodes*, Commission inter-IREM Lycées techniques, brochure nº 112, 2001.

Groupe Probabilités & statistique de l'IREM de Besançon : *Lois continues, tests d'adéquation, une approche pour non spécialistes*, Presses universitaires de Franche-Comté, 2005.

Saporta, G. : *Probabilités, Analyse des données et Statistique,* Technip, Paris, 1990.

Traitement de l'exercice 3 sur Excel

	A	B	C	D
	Nombres aléatoires	SB: Chiffres équirépartis	SC: répartition biaisée	Biais
1	=ALEA()	=ENT(10*A2)	=SI(ET(B2/2=ENT(B2/2);(ENT(1000*A2-100*B2>D$2)));B2+1;B2)	
2	0,522488616	5	5	65
3	0,679476752	6	7	
	etc	etc	etc	

	E	F	G	H	I	
	Taille n	Effectifs B	Effectifs C	Distances	Tests	
2	100					
3				d_B^2	Test précis	
4	seuil a			5,6	SI(H4>E11;1;0)	
5	0,05				Test simplifié	
6					SI(H4>2*E8;1;0)	
7	Modalités k					
8	10			d_C^2	Test précis	
9				13,2	SI(H9>E11;1;0)	
10	valeur critique				Test simplifié	
	KHIDEUX.INVERSE(E5;E8-1)				SI(H9>2*E8;1;0)	
11	16,91897762					
12		=FREQUENCE(B2:INDIRECT(ADRESSE(E2+1;2;1;1;"Feuil2"));E14:E23)				
13	Modalités		=FREQUENCE(C2:INDIRECT(ADRESSE(E2+1;3;1;1;"Feuil2"));E14:E23)			
14	0	14	9			
15	1	12	17			
16	2	8	4			
17	3	11	15			
18	4	9	8			
19	5	11	12			
20	6	10	7			
21	7	7	10			
22	8	12	8			
23	9	6	10			