Adéquation à une loi équirépartie

Une expérience aléatoire conduisant à k issues x_i dont les probabilités respectives sont p_i est réalisée n fois. On note n_i la variable aléatoire égale au nombre de réalisations de l'issue x_i sur les n expériences. (on a donc n_i [Acs]> B(n,p_i))

Test du χ^2 (Khi-Deux)
Ce test est basé sur le calcul des distances entre les valeurs observées et les valeurs théoriques. Il est utilisable uniquement pour les lois discrètes. Si la loi est continue, on se ramène à une loi discrète en regroupant les données en classes de la forme $]x_i,x_{i+1}]$. Dans chaque classe, on doit obtenir un nombre théorique $np_i \geqslant 5$; l'intervalle de cette classe n'étant pas forcément régulier.

Pour mesurer ces écarts, on considère la variable aléatoire $D^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

On montre que l'espérance mathématique de D^2 est indépendante de n et des p_i : $E(D^2) = k - 1$. De plus, on montre que la loi de D^2 ne dépend pas de n dès que n est assez grand ; en pratique pour $n \geqslant 30$, mais ne dépend que du nombre k d'issues. Cette loi est proche de celle d'une variable unique notée χ^2 dont les vingtiles v_α sont donnés par la table ci-après. lorsque la loi théorique est connue, $E(D^2) = k - 1$ est appelé "nombre de degrés de liberté" de cette variable.
On convient, pour une expérience aléatoire, de rejeter la "régularité" de façon significative au seuil de 10% (ou de 5%) lorsque la valeur observée de D^2 dépasse $v_{18} = d_9$ (ou v_{19} pour 5%) :
Soit d^2 la valeur observée de D^2,
si $d^2 < v_{19}$, au seuil de 5%, on ne peut pas rejeter l'hypothèse initiale sur le modèle (équiréparti ?)
si $d^2 > d_9$, au seuil de 10%, on décide de rejeter l'hypothèse de régularité

k - 1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
v_{18}=d_9	2,71	4,61	6,25	7,78	9,24	10,64	12,02	13,36	14,68	15,99	17,27	18,55	19,81	21,06	22,31
v_{19}	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92	18,31	19,67	21,03	22,36	23,68	24,00

k - 1	16	17	18	19	20	25	30	40	50	60	70	80	90	100
v_{18}=d_9	23,54	24,77	25,99	27,20	28,41	34,38	40,26	51,80	63,17	74,40	85,53	96,58	107,57	118,50
v_{19}	26,30	27,59	28,87	30,14	31,41	37,65	43,77	55,76	67,50	79,08	90,53	101,88	113,15	124,34

Utilisation de l'écart quadratique réduit

Il est défini par la distance $d^2 = \sum_{i=1}^{k} \frac{(f_i - p_i)^2}{p_i}$ où f_i sont les fréquences observées. On a donc $D^2 = n.d^2$

Comme pour le test du χ^2, si les valeurs observées sont proches les unes des autres, d^2 sera petit et on considérera qu'il y a adéquation entre l'expérience et la loi de probabilité équirépartie. Le problème est de savoir à partir de quelle borne, les valeurs de d^2 peuvent être considérées comme assez petites pour que le test soit jugé positif.
On réalise n fois une simulation d'un certain nombre de fois l'expérience.
Lorsque n est "assez grand", le neuvième décile d'une simulation est constant et indépendant des séries.
On convient alors qu'il n'y a pas d'adéquation entre l'observation et la théorie, au risque de 10% si $d^2 > d_9$.
Plus précisément, si la théorie est vraie, il y aura 10 chances sur 100 de refuser une adéquation valide.

Une autre variante dans le cas d'une loi théorique équirépartie

Cette méthode ne fonctionne que si la loi théorique suit une loi équirépartie., On suppose toujours une loi ayant k issues ; chaque p_i est donc égale à $\frac{1}{k}$.

On note $d'^2 = \sum_{i=1}^{k} \left(f_i - \frac{1}{k}\right)^2$, On a donc $D^2 = nk.d'^2$ ou $d^2 = k.d'^2$

On considère qu'il est très rare (moins de 5% des échantillons possibles ou au niveau de confiance de 95%) que d'^2 dépasse $\frac{2}{n}$.

Ou encore, en considérant les effectifs plutôt que les fréquences, si $\sum_{i=1}^{k} \left(n_i - \frac{n}{k}\right)^2$ dépasse $2n$.

<u>Remarques</u> :

(i) Pour $k = 2$, épreuve à deux issues équiprobables, on retrouve l'intervalle de dispersion vu en seconde à l'occasion des sondages :

$$d'^2 = \left(f_1 - \frac{1}{2}\right)^2 + \left(f_2 - \frac{1}{2}\right)^2 = \left(f_1 - \frac{1}{2}\right)^2 + \left(1 - f_1 - \frac{1}{2}\right)^2 = 2\left(f_1 - \frac{1}{2}\right)^2$$

La condition $d'^2 < \frac{2}{n}$ est donc équivalente à la condition $\left(f_1 - \frac{1}{2}\right)^2 < \frac{2}{n}$ *ie* $\left|f_1 - \frac{1}{2}\right| < \frac{1}{\sqrt{n}}$

(ii) Pour un autre niveau de confiance, il faut changer $\frac{2}{n}$ par

$\frac{1{,}6}{n}$ pour un niveau de confiance de 90%

$\frac{3{,}4}{n}$ pour un niveau de confiance de 99%

(iii) Les valeurs de d'^2 obtenues dans chaque simulation diminuent quand on augmente le nombre m de lancers de chaque série car les fluctuations d'échantillonnage sont moindres pour les grands échantillons. Par contre, on admet que les valeurs de $m.d'^2$ sont du même ordre de grandeur, quelle que soit la valeur de m, pourvu que m soit assez grand, disons pour $m \geqslant 100$.