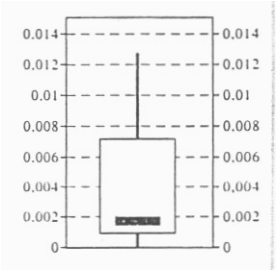
STATISTIQUE

I Poser le problème de l'adéquation de données

Enoncé 1 :

On veut tester si une pièce de monnaie est truquée ou non. Pour cela, on la lance 100 fois. On obtient 59 fois "pile"et 41 fois "face". Au seuil de risque 10 peut-on dire que cette pièce est truquée ?
On utilisera les résultats de la simulation de cette expérience répétée 1 000 fois, pour laquelle on a calculé le nombre d^2 somme des carrés des écarts entre les fréquences observées et les fréquences théoriques.
On donne ci-contre le diagramme en boîte de la série statistique des valeurs de d^2.



Solution :

Méthode : Pour tester une hypothèse au seuil de risque 10%, utiliser la neuvième décile de la série des valeurs simulées.
La fréquence observée des "pile" est 0,59 et celle observée des "face" est 0,41. Comme la loi uniforme sur $\Omega = (P ; F)$ est telle que $P\{P\} = P\{F\} = 0,5$ le calcul de d^2 associé aux données de l'expérience s'écrit :

$$d^2 = (0,59 - 0,5)^2 + (0,41 - 0,5)^2 = 0,0162.$$

Le neuvième décile de la série statistique des nombres d^2 obtenus par simulation est environ 0,013.
Comme 0,0162 est supérieur à D_9, on peut considérer, au risque 10 %, que cette pièce de monnaie est truquée.

Enoncé 2 :

Dans une maternité, on a noté pendant un an l'heure de chaque naissance. Les nombres de naissances entre 0h et 1h, entre 1h et 2h,…, sont respectivement 96, 126, 130, 125, 124, 129, 115, 89, 118, 97, 95, 108, 98, 97, 109, 95, 115, 108, 90, 104, 103, 112, 113, 128. Tester au seuil de risque 10% si une naissance se produit avec la même probabilité dans l'une des 24 heures. Au cours de 2000 simulations de cette expérience, on a calculé le nombre d^2, somme des carrés des écarts entre les fréquences observées et les fréquences théoriques. Voici les résultats pour la série statistique des valeurs de $10^4\ d^2$

Minimum	D_1	Q_1	Médiane	Q_3	D_9	Maximum
0.6	16.9	23.2	25.8	32.1	36.5	61

Solution :

Méthode : Pour définir la probabilité uniforme, déterminer le nombre q d'issues ; elle est alors égale à 1/q.
Les fréquences observées pour chaque intervalle d'une heure de la journée sont : 0,0366 – 0,0.480 – 0,0495 – 0,476 - 0,0473 - 0,0492 - 0,0438 - 0,0339 - 0,0450 - 0,0370 - 0,0362 - 0,0412 - 0,0373 – 0,0370 - 0,0415 - 0,0362 - 0,0438 - 0,0412 - 0,0343 - 0,0396 - 0,0393 - 0,0427 - 0,0431 - 0,0488. La loi uniforme sur {1 ; 2 : 3 ; … ; 24} est telle que la probabilité de chaque événement élémentaire est 1/24. On calcule alors la valeur de d^2 issue de l'observation :

$$d^2 = \left(0.0366 - \frac{1}{24}\right)^2 + \left(0.0480 - \frac{1}{24}\right)^2 + \ldots + \left(0.0488 - \frac{1}{24}\right)^2$$

On trouve $d^2 \approx 0.0006$, soit environ 6.10^{-4}
Cette valeur de d^2 est inférieure au neuvième décile ($36,5.\ 10^{-4}$) de la série des valeurs simulées. Ainsi, au seuil 10%, on peut dire q'une naissance se produit avec la même probabilité toutes les 24 heures dans cette maternité

II Adéquation de données à une loi équirépartie

Un joueur veut vérifier si le dé qu'il possède est "normal", c'est-à-dire bien équilibré. On sait que, dans ce cas-là, la loi de probabilité associée est la loi uniforme : P{1} = P{2} = P{3} = P{4} = P{5} = P{6} = 1/6
Pour cela, le joueur lance 200 fois le dé et note les résultats obtenus :

x_i	1	2	3	4	5	6
n_i	31	38	40	32	28	31
f_i	0.155	0.190	0.200	0.160	0.140	0.155

Pour savoir si la distribution de fréquences obtenue est « proche » de la loi uniforme, on calcule la quantité suivante, qui prend en compte l'écart existant entre chaque fréquence trouvée et la probabilité théorique attendue:

$$d^2 = \left(0.155 - \frac{1}{6}\right)^2 + \left(0.190 - \frac{1}{6}\right)^2 + ... + \left(0.155 - \frac{1}{6}\right)^2 \approx 0.00268$$

Notation : On note la quantité d^2 car son calcule est celui du carré d'une distance.

Mais rien ne permet de dire pour l'instant si cette quantité trouvée est « petite » ou « grande ». En effet, elle est soumise à la fluctuation d'échantillonnage, puisque sa valeur varie d'une série de lancers à l'autre. On va donc étudier cette fluctuation d'échantillonnage pour convenir d'un seuil entre « petite » et « grande » valeur de d^2 lorsqu'on lance 200 fois un dé. Pour cela, on génère des séries de 200 chiffres au hasard pris dans {1 ; 2 ; 3 ; 4 ; 5 ; 6}. Les résultats trouvés pour le nombre d^2 à partir de 1 000 simulations sont résumés par le tableau suivant :

Minimum	D_1	Q_1	Médiane	Q_3	D_9	Maximum
0.00363	0.00138	0.00233	0,00363	0.00555	0.00789	0.01658

Technique : Q_1 et Q_3 sont le premier et le troisième quartile et D_1 et D_9 sont le premier et le neuvième décile de la série.

Le **neuvième décile** de la série des valeurs simulées de d^2 est 0,00789.
Cela signifie que 90% des valeurs de d^2 obtenues au cours de ces 1 000 simulations sont dans l'intervalle [0 ; 0.00789]. Comme la valeur observée de d^2 est inférieure à cette valeur seuil de 0,00789, on peut convenir que le dé est équilibré avec un risque de 10%.
En effet, en utilisant celle méthode sur les données simulées, on se serait trompé dans 10% des cas. On dit que l'on a un **seuil de confiance** de 90 %.

Propriété :

Soit une épreuve conduisant aux issues a_1, a_2, …,a_q.
Expérimentalement, si on répète n fois cette épreuve ($n \geq 100$), on obtient les fréquences f_1, f_2, …, f_q pour chacune des issues. Pour vérifier l'adéquation de ces données à la **loi équirépartie** sur $\{a_1, a_2, ..., a_q\}$, on calcule le nombre :

$$d^2 = \sum_{i=1}^{q} \left(f_i - \frac{1}{q}\right)^2 .$$

Technique : Le processus décrit ici est un cas particulier simplifié d'un processus beaucoup plus général et très utilisé en statistiques : le test du χ^2 (khi-deux).

La réalisation d'un grand nombre de simulations de cette épreuve conduit pour la variable d^2 à une série statistique de neuvième décile D_9.
Si $d^2 \leq D_9$ alors on dira que les données sont compatibles avec le modèle de la loi uniforme au seuil de risque 10 %.
Si $d^2 \geq D_9$ alors on dira que les données ne sont pas compatibles avec ce modèle au seuil de risque 10 %.