ADÉQUATION À UNE LOI ÉQUIRÉPARTIE

Activités informatiques sous tableur

Objectif de l'activité :

Comment, à partir d'un échantillon de données expérimentales liées à un phénomène aléatoire, décider, par une argumentation de nature statistique, qu'un modèle mathématique est en **adéquation** avec la réalité?

1 Introduction :

Lorsqu'on joue avec un dé, on souhaite savoir s'il est bien équilibré, c'est à dire si chaque face a autant de "chances" de sortir; il y a là un modèle mathématique sous jacent : "l'équiprobabilité". En pratique, même pour un dé dont on sait qu'il est équilibré, le tableau des fréquences ne correspond pas (sauf cas très exceptionnel) à la loi de probabilité (il lui correspond si le nombre de lancer est infini : c'est la loi des grands nombres).

Pour tester si le modèle d'équiprobabilité est effectivement adapté au dé utilisé, on réalise l'expérience suivante : on joue 200 fois avec le dé, on construit une statistique de cet échantillon, on en fait le traitement (représentation graphique, paramètres de position et de dispersion) et on compare avec les données théoriques.

Exemple : On a trouvé sur une série de 200 lancers du dé la répartition suivante :

Faces	1	2	3	4	5	6
Effectifs	45	42	32	29	24	28

Peut-on raisonnablement estimer que le dé est bien équilibré ou non ?

On ne peut être sûr de rien mais on peut, d'un point de vue statistique, avoir une idée valide de la réponse grâce aux simulations.

Etude de cette série de lancers : ouvrir le fichier *adeq_elv.xls* , à la page *dé.*

Compléter : colonne C les fréquences f_i, colonne D les produits $f_i \times x_i$, en cellule D10 la moyenne de cette série; la colonne E correspond aux probabilités $\frac{1}{6}$ (équiprobabilité) et la colonne F aux produits $p_i \times x_i$; la moyenne attendue avec un dé parfait est en cellule D10.

On observe que les deux moyennes sont différentes.

Pour savoir si cette différence est significative, on effectue une **simulation** d'une loi équirépartie sur {1; 2; 3; 4; 5; 6} à l'aide du tableur : l'ordinateur permet, grâce à la formule Ent(6 * alea( ) + 1) d'obtenir *au hasard* un nombre entier compris entre 1 et 6, ce qui simule ainsi un dé parfait. On compare ensuite les données expérimentales collectées précédemment avec les résultats simulés par l'ordinateur.

2 Simulation :

1ère étape : simulation de 500 échantillons de même taille que la série expérimentale de ce "dé parfait" (taille 200).

Ouvrir la page *séries multiples* : écrire dans la cellule N7 la formule **=Ent(6*alea()+1)**, puis faire une recopie automatique jusqu'en cellule HE7 pour simuler les 200 lancers sur une ligne; recopier ensuite automatiquement cette ligne de 2OO cellules jusqu'à la ligne 506 pour obtenir, en colonne, 500 échantillons de 200 lancers.

2ème étape : calcul de la fréquence d'apparition de chacun des nombres 1 à 6 dans les 500 échantillons.

Ecrire en cellule B7 la formule **=NB.SI($N7:$HE7;B$6)/200** (la fonction NB.SI permet de compter le nombre d'occurrence de B$6 dans la plage $N7:$HE7); faire une recopie automatique jusqu'en G7 pour calculer toutes les fréquences d'apparition des nombres 1 à 6 sur le premier échantillon de 200 lancers.

La moyenne des points du dé parfait sur ce premier échantillon est donnée en cellule I7 par la formule **=B$6*B7+C$6*C7+D$6*D7+E$6*E7+F$6*F7+G$6*G7**, et le carré de l'écart à la moyenne théorique de 3,5 donnée en cellule J7 par la formule **=(I7-3,5)^2**.

Recopier ensuite automatiquement la ligne de B7 à J7 jusqu'à la ligne 506 pour effectuer les mêmes calculs sur les 500 échantillons.

C'est en utilisant les écarts à la moyenne que l'on va d'abord tester le dé.

3 Étude de la moyenne :

Une étude statistique des moyennes d'une série de 500 échantillons va permettre de mieux cerner les **fluctuations d'échantillonnage** des échantillons de taille 200.

Ouvrir la page *étude d'une série (moyenne).* Les données d'une série obtenue en page précédente ont été recopiées et vont être étudiées.

Tout d'abord, on fait un regroupement des 500 moyennes en classes d'amplitude 0,1 ou 0,05 : les fréquences cumulées croissantes de chaque classe sont obtenues en écrivant la formule **=FREQUENCE(I:I;K7:L7)/500** en cellule M7, recopiée jusqu'en M24; pour les fréquences, écrire **=M7** en cellule N7 et **=M8–M7** en N8, à recopier jusqu'en N24; pour les fréquences cumulées décroissantes, écrire **1** en cellule O7 et **=O7–N7** en cellule O8 à recopier jusqu'en O24.

Représenter la série des fréquences par un histogramme.

On constate statistiquement que la répartition des moyennes donne un histogramme en forme de cloche : 5 % des moyennes sont inférieures à 3,3 et 5 % sont supérieures à 3,7; donc 90 % des moyennes sont entre 3,3 et 3.7.

On adopte la règle de décision suivante : si la moyenne observée dans l'échantillon obtenu avec le dé n'est pas dans cet intervalle, on décide de dire qu'il y a "peu de chance" que le dé soit parfait.

Pour l'instant le dé a "moins de 10 % de chance" d'être parfait puisque la moyenne trouvée 3,145 n'est pas dans l'intervalle [3,3 ; 3,7].

On va affiner l'étude en s'intéressant aux fréquences seules.

4 Étude de la distance entre les fréquences observées et la fréquence théorique :

Revenir à la page *dé* : sur la série expérimentale des 200 lancers du dé, on calcule en colonne G les nombres : $\left(f_i - \frac{1}{6}\right)^2$, carrés des écarts à la moyenne théorique, et en cellule G9, la somme des cellules G3 à G8 donnant la "distance observée" : $d_{obs}^2 = \sum_{i=1}^{i=6} \left(f_i - \frac{1}{6}\right)^2$.

La distance observée 0,008 683 n'est pas nulle. Pour savoir si cette distance est la conséquence d'un dé mal équilibré, on utilise de nouveau la simulation : l'étude statistique des distan ces calculées à partir des 500 échantillons simulés de taille 200 permet de mieux cerner les fluctuations d'échantillonnage sur les distances.

Se placer à la page *étude d'une série (distance).*

Ecrire **=(B5-1/6)^2+(C5-1/6)^2+(D5-1/6)^2+(E5-1/6)^2+(F5-1/6)^2+(G5-1/6)^2** en cellule J5, recopier jusqu'en J504.

Recopier la colonne J en colonne K : pour cela, sélectionner les cellules J5 à J504 (il est conseillé de commencer par J504 et remonter à J5…), se placer en K5 et faire un *collage spécial* en choisissant *valeurs*; faire un tri croissant de la colonne K; le 9ème décile s'obtient par la formule **=CENTILE(K5:K504;0,9)** (en cellule T2).

Représenter graphiquement par un nuage de points la série triée.

On constate statistiquement que 90 % des distances sont inférieures à 0,008 (car le 9ème décile vaut 0,008).

On adopte la règle de décision suivante : si la distance observée dans l'échantillon fabriqué à partir du dé est plus grande que le 9ème décile 0,008, on décide de dire qu'il y a "peu de chance" que le dé soit parfait.

5 Conclusion :

Avec ce critère, le dé a moins de 10 % de chance d'être parfait et sera déclaré non conforme au modèle.

Remarque : Le seuil choisi ici en tant que critère de décision est de 10%. En procédant ainsi, on risque de rejeter à tort l'hypothèse d'équiprobabilité dans au maximum 10% des cas. Avec les mêmes valeurs expérimentales et un seuil différent, on peut tester et accepter l'hypothèse. On ne démontre donc pas que le dé est équilibré.

Critères de décision :

- Si d_{obs}^2 ☝ D_9, on **accepte** le modèle d'équirépartition, c'est-à-dire que l'on considère que les résultats observés sont compatibles avec l'hypothèse que le dé est bien équilibré.
- Si d_{obs}^2 ☜ D_9, on rejette l'hypothèse que le dé est équilibré **au risque d'erreur de 10%.**

5 Adéquation à une loi équirépartie :

Pour décider si le modèle d'équirépartition peut être adopté pour une expérience aléatoire ayant k issues (6 dans le cas du lancer de dé), on réalise cette expérience n fois, avec n assez grand (supérieur à 30) et on compare ces résultats avec une situation identique dans laquelle l'équirépartition est établie.

On se base sur l'étude des distances : on calcule la distance observée $d_{obs}^2 = \sum_{i=1}^{i=k} \left(f_i - \frac{1}{k}\right)^2$ entre les fréquences obtenues et la fréquence théorique $\frac{1}{k}$.

Ensuite, on effectue un grand nombre N de simulations de ces n épreuves dans les conditions vues précédemment (au moins 100 simulations de n épreuves), on étudie la série des distances obtenues avec la même formule (carrés des écarts avec la valeur théorique), et on choisit un **seuil de référence** selon la précision souhaitée :

- Si l'on choisit D9 (cas le plus fréquent), cela signifie que :

lorsque d_{obs}^2 ☜ D_9 **on rejette l'hypothèse d'équirépartition au risque d'erreur de 10%**
lorsque d_{obs}^2 ☝ D_9 on peut accepter l'hypothèse d'équirépartition, risque de commettre une erreur, donc sans certitude, ni même évaluation du risque d'erreur.

- Si l'on choisit le 95ème centile : le risque d'erreur en cas de rejet est de 5%.

Remarques :

1) Par ces simulations, on ne prouve pas que le dé est équilibré : on fait une "étude comparative" qui donne une idée plus ou moins valide de la réponse.

2) Plus le seuil est élevé, moins on prend de risque à rejeter l'hypothèse d'équirépartition : si l'on choisit D9 ou le 95ème centile comme seuil, le dé est considéré comme non équilibré avec une marge d'erreur de 10%.ou 5%. Une marge d'erreur de 1% correspond au 99ème centile : cela signifie qu'on considère que seulement 1% des résultats de la simulation sont marginaux. Ici, si on choisit le seuil de 1%, comme le 99ème centile vaut 0,014, le dé étudié est considéré comme équilibré : on ne prend donc pas le risque de dire que le dé est déséquilibré, mais on n'est pas sûr pour autant qu'il soit équilibré! C'est là tout le paradoxe.

3) Dans la pratique, on n'étudie pas les moyennes, seulement les distances entre fréquences observées et fréquences théoriques (dont dépendent les moyennes) ; de plus, ces écarts étant faibles, on utilise souvent $n \times d_{obs}^2$ pour l'étude.