Approche empirique du test χ^2 d'ajustement

Alain Stucki, Lycée cantonal de Porrentruy

Introduction

En lisant des rapports, on rencontre souvent des raisonnements du style : « le premier groupe est meilleur que le deuxième avec un taux de réussite de 8% supérieur ».

Il n'est pas toujours simple de faire comprendre aux auteurs de telles conclusions que les différences des pourcentages calculés ne sont peut-être pas significatives, et qu'il est indispensable d'utiliser des méthodes statistiques et probabilistes pour mesurer le degré de confiance que l'on peut déduire de ces différences.

Nous nous sommes trouvés face à un problème de ce genre lorsque, dans le cadre de leur travail de maturité, deux élèves ont voulu mettre en évidence que le groupe sanguin des lycéens influence le choix de leur option spécifique. Après avoir récolté leurs données, ces deux élèves sont tombés dans le piège de la comparaison de pourcentages.

C'est la raison pour laquelle nous avons imaginé ensemble une introduction aux tests d'hypothèse qui ne nécessite pas de connaissance préalable en statistiques et en calcul des probabilités. Cet article présente une approche empirique du test χ^2 d'ajustement (aussi appelé test χ^2 d'adéquation). Il est simple et applicable dans de nombreuses situations où les données sont regroupées en catégories.

Le dé est-il bien équilibré ?

On lance 400 fois un dé tétraédrique dont les faces sont numérotées de 1 à 4 (le choix d'un dé tétraédrique est en relation avec les quatre groupes sanguins). Si le dé est bien équilibré, on s'attend à environ 100 apparitions de chaque face.

Les données expérimentales ainsi que les premiers calculs que nous inspirent ces données sont présentés dans les tableaux ci-dessous. Ce sont de tels tableaux qui sont souvent à l'origine de conclusions hâtives.

Numéro de la face	1	2	3	4
Fréquences observées o_i	77 (19.25%)	102 (25.50%)	113 (28.25%)	108 (27.00%)
Fréquences espérées e_i	100 (25%)	100 (25%)	100 (25%)	100 (25%)

TAB. 1 – Fréquences observées et fréquences espérées

Numéro de la face	1	2	3	4
$\dfrac{o_i - e_i}{e_i} \times 100$	$\dfrac{77-100}{100} \times 100 = -23\%$	2%	13%	8%

TAB. 2 – Fréquences relatives en pourcentages

On remarque dans le tableau 1 que le dé a montré seulement 19.25% de faces marquées 1 contre les 25% espérés, ou encore que le nombre d'apparitions de la face marquée 3 dépasse de 33 celui de la face marquée 1.

Dans le tableau 2, on peut constater que le nombre observé de faces marquées 1 est de 23% inférieur à la valeur espérée ou encore que le nombre observé de faces marquées 3 est supérieur de 13% à la valeur théorique.

Les constatations précédentes font apparaître des valeurs troublantes. Est-ce suffisant pour conclure que le dé n'est pas bien équilibré ? Autrement dit, peut-on suspecter que le comportement de ce dé n'est pas dû au hasard seul ?

Pour évaluer si la différence entre les fréquences observées et celles espérées est significative, nous allons déterminer une mesure du « degré de différence » entre les deux séries.

Pour ce faire, calculons la somme des carrés relatifs des différences que l'on note χ^2.

$$\chi^2 = \frac{(77-100)^2}{100} + \frac{(102-100)^2}{100} + \frac{(113-100)^2}{100} + \frac{(108-100)^2}{100} = 7.66$$

Il est évident que plus cette somme est grande, plus les fréquences observées s'éloignent des fréquences espérées. Un dé idéal donnerait $\chi^2 = 0$.

Comment estimer raisonnablement si la valeur $\chi^2 = 7.66$ est due aux fluctuations naturelles des résultats obtenus avec un dé bien équilibré ou à un mauvais équilibrage du dé ayant servi à l'expérience ?

Pour répondre à cette question, nous allons calculer expérimentalement la fréquence avec laquelle une série obtenue au hasard avec un dé normal donne un χ^2 égal ou supérieur à 7.66.

Modélisation

Nous avons utilisé *Mathematica* et son générateur de nombres (pseudo-)aléatoires pour modéliser l'expérience « lancer $100k$ fois un dé bien équilibré à k faces ». Les fréquences espérées sont ainsi toutes égales à 100. A l'issue d'une réalisation de l'expérience, le χ^2 est calculé. On obtient alors une distribution empirique des χ^2.

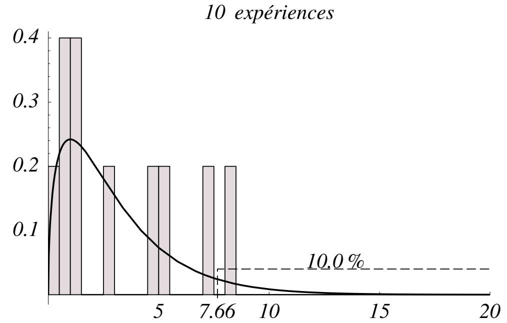
Nous avons écrit le module **`chicarre[deglib_,nbexp_,valeur_]`** qui dessine l'histogramme des fréquences des χ^2 obtenus en réalisant un nombre **`nbexp`** d'expériences et qui calcule le pourcentage de χ^2 supérieurs à la valeur **`valeur`** que nous désirons tester. La distribution théorique de *Pearson*[1] correspondant au nombre **`deglib`** de degrés de liberté $(k-1)$ est également représentée. Dans l'exemple qui nous intéresse, le nombre de degré de liberté vaut 3 et la valeur testée est égale à 7.66.

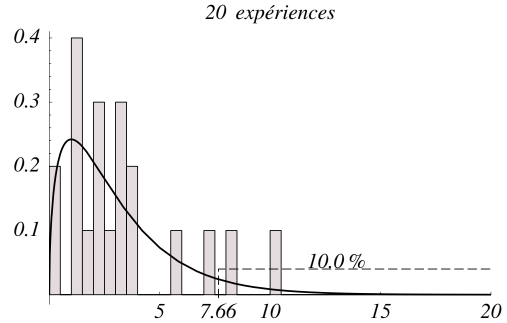
Ce module est en annexe de la version électronique du présent article, disponible sur le site de la Commission Romande de Mathématique (`http://www.sspmp.ch/crm/avis.htm`).

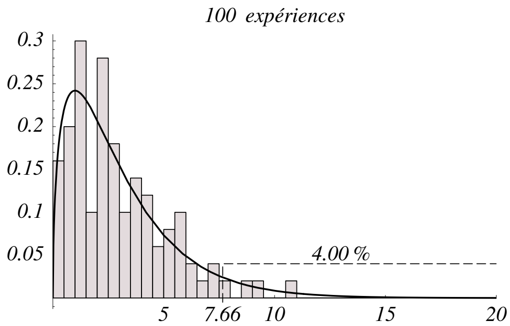
Les élèves jouent volontiers avec le module. Il est cependant important de les inviter à débuter avec un petit nombre d'expériences (au début, la répétition d'un seul tirage permet de mettre en évidence les fluctuations naturelles). Par la suite, les gros tirages peuvent être entrepris. C'est l'occasion de réaliser qu'un grand nombre d'expériences peut être modélisé (ici par la distribution théorique de Pearson), alors que les petits tirages sont imprévisibles.

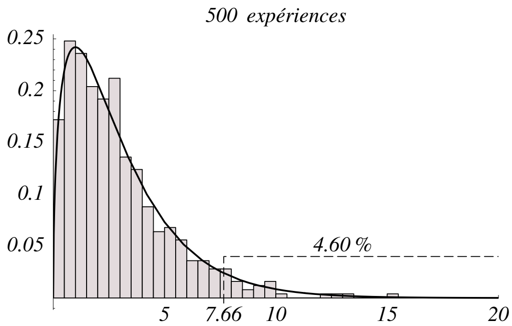
Les figures qui suivent montrent quelques exemples de graphiques obtenus pour la valeur de χ^2 qui nous intéresse.

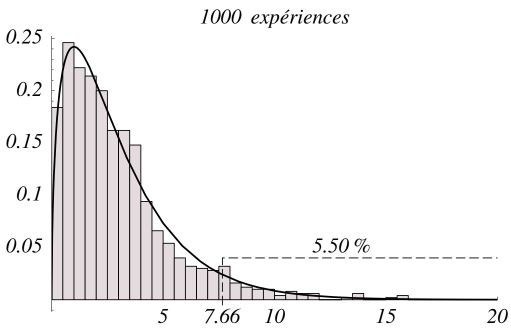
[1] Karl Pearson (1857-1936), mathématicien, physicien et historien anglais, a mis en valeur cette distribution qui avait déjà été étudiée par l'astronome et géodésien allemand Friedrich Robert Helmert (1843-1917).

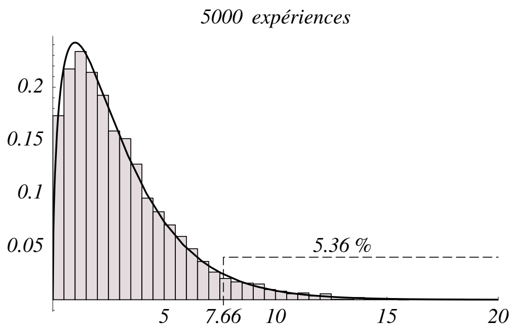












La dernière distribution (5 000 expériences) nous indique que pour un dé bien équilibré, une série de 400 lancers donne un χ^2 supérieur ou égal à 7.66 dans 5.36% des cas (pourcentage de l'aire de l'histogramme située à droite de 7.66).

Ainsi, en rejetant l'hypothèse que le dé ayant servi à notre première expérience est bien équilibré, on encourt le risque de se tromper avec une probabilité de 5.36% (le pourcentage possible dû uniquement au hasard). La valeur donnée par la distribution théorique de *Pearson* est 5.34%. Ce risque est dit de *première espèce*.

Conventionnellement, on fixe un seuil de confiance à 5%. Dans notre cas, on ne peut pas conclure que le dé n'est pas bien équilibré.

Nous acceptons donc que le dé est normal avec un autre risque : celui d'accepter une hypothèse fausse (que le dé est bien équilibré alors qu'il ne l'est pas). Ce risque, dit de *deuxième espèce*, ne peut pas être calculé.

Le protocole du test χ^2 d'ajustement

Ce test permet de déterminer si une population répartie en catégories suit une distribution connue (par exemple la distribution d'une population de référence). Le protocole est le suivant.

1. On note H_0 l'hypothèse, dite nulle, que la population répartie en k catégories suit une distribution connue, et on choisit un *seuil* α (le plus souvent α est fixé à 5%). Ce seuil représente le risque, en n'acceptant pas H_0, d'avoir rejeté une hypothèse qui était exacte.
2. Pour $1 \leq i \leq k$, on note o_i l'effectif observé expérimentalement et e_i l'effectif espéré (celui correspondant à la distribution connue). Le test est applicable si tous les e_i sont supérieurs à 5, et on cacule alors la grandeur
$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \frac{(o_1 - e_1)^2}{e_1} + \frac{(o_2 - e_2)^2}{e_2} + \cdots + \frac{(o_k - e_k)^2}{e_k}.$$
3. On rejette l'hypothèse H_0 au seuil α si $\chi^2 \geq \chi^2_\alpha$, la valeur de χ^2_α étant lue dans une table du test du χ^2 à $k - 1$ degrés de liberté. Sinon on admet H_0.

Remarques

1. Les valeurs qui doivent être lues dans la table correspondent aux valeurs calculées à l'aide de la distribution théorique de *Pearson* à $k - 1$ degrés de liberté. Pour 3 degrés de liberté, la courbe de la distribution théorique est celle qui est représentée sur les graphiques précédents. Nous espérons que les élèves qui ont effectué les simulations et qui ont eu sous les yeux l'histogramme empirique ont pu donner un sens aux valeurs tabulées.
2. Le seuil α est la probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie. La probabilité β d'accepter l'hypothèse H_0 alors qu'elle est fausse ne peut en général pas être calculée. Le tableau suivant résume les erreurs dans le processus de décision.

Réalité Décision	H_0 est vraie	H_0 est fausse
H_0 est rejetée	Risque de première espèce $p = \alpha$	Décision correcte $p = 1 - \beta$
H_0 est acceptée	Décision correcte $p = 1 - \alpha$	Risque de deuxième espèce $p = \beta$

Tab. 3 – Erreurs de décision

   Notons que de diminuer le risque de première espèce en fixant par exemple $\alpha = 1\%$ a pour conséquence d'augmenter le risque de deuxième espèce. Le niveau de α dépend essentiellement de l'importance relative des erreurs de première et de deuxième espèce dans le cadre de l'étude que l'on réalise. Dans la pratique, on admet en général $\alpha = 5\%$.
3. Le fait d'admettre H_0 ne signifie pas que nous avons démontré que cette hypothèse est exacte, mais seulement que nous n'avons pas de raisons suffisantes de la rejeter ou encore que nous n'avons pas rassemblé suffisamment de données pour conclure. Le test est d'ailleurs souvent utilisé en vue de rejeter l'hypothèse nulle, donc pour mettre en évidence une différence.

Exemple

Un chocolatier a créé trois nouveaux pralinés (A, B et C). Pour des raisons économiques, un seul praliné sera commercialisé. Il décide donc de tester ses nouvelles friandises sur 90 personnes afin de déterminer si l'un des nouveaux mélanges de chocolat est plus appécié que les deux autres.

Chaque personne a goûté les trois pralinés et a sélectionné celui qu'elle préfère. Le test du χ^2 est appliqué.

H_0 : Les trois variétés sont également appréciées.

	A	B	C	Total
Fréquences observées o_i	29	40	21	90
Fréquences espérées e_i	30	30	30	90

$$\chi^2 = \frac{(29-30)^2}{30} + \frac{(40-30)^2}{30} + \frac{(21-30)^2}{30} \simeq 6.07.$$

La table nous donne $\chi^2_{0.05} = 5.99$. Comme $6.07 > 5.99$, nous rejetons H_0 au seuil $\alpha = 5\%$. Les trois variétés ne sont pas également appréciées, et selon les résultats le praliné B semble plus estimé que les deux autres.

L'exemple de mes élèves

Le groupe sanguin d'un étudiant influence-t-il son choix d'option spécifique ? Prenons le cas des étudiants ayant choisi une option scientifique. Mes élèves ont considéré comme scientifiques les options « biologie-chimie » et « physique-applications des mathématiques ». Ils ont interrogé 115 étudiants de ces sections. La population de référence est la population suisse.

H_0 : La répartition des étudiants scientifiques est identique à la répartition de la population suisse.

	A	B	AB	O	Total
Fréquences observées o_i	48	6	5	56	115
Répartition de la population suisse	47%	8%	4%	41%	100%
Fréquences espérées e_i	54.05	9.20	4.60	47.15	115

$$\chi^2 = \frac{(48-54.05)^2}{54.05} + \frac{(6-9.2)^2}{9.2} + \frac{(5-4.6)^2}{4.6} + \frac{(56-47.15)^2}{47.15} \simeq 3.49.$$

La table du χ^2 à 3 degrés de liberté nous donne $\chi^2_{0.05} = 7.82$. Comme $3.49 < 7.82$, nous n'avons pas de raison de penser que la répartition des élèves scientifiques selon leur groupe sanguin est différente de celle de la population suisse.

L'étude que ces élèves ont menée est inspirée de l'ouvrage « Le secret de votre groupe sanguin, clé de la personnalité et du destin » de Jean-Louis Degaudenzi aux éditions Filipacchi. En lisant ce livre, on apprend qu'au Japon le groupe sanguin est un critère déterminant dans le recrutement de personnel. Autrement dit, on ne s'improvise pas danseur étoile au pays du soleil levant, c'est écrit dans le groupe sanguin.