LES TRUITES DE PONDICHERY

Un problème d'adéquation au baccalauréat

Bernard KOCH
et le Groupe Statistiques [1]
Irem de Strasbourg

Plusieurs sujets du Bac 2003 comportaient des exercices de probabilités comme celui qui figure en tête du sujet de ES de Pondichery (commun à tous les candidats). Il s'agit de tester l'adéquation d'une distribution de fréquences observées à un modèle théorique équiprobable.

Le principe d'un test d'adéquation

Pour tester si une expérience aléatoire avec, par exemple, trois issues possibles (notées ici 1,2 et 3), est conforme au modèle où les trois valeurs sont équiprobables on l'observe n fois. On dispose alors du tableau des *effectifs observés* :

Valeurs	1	2	3	Total
Effectifs	e_1	e_2	e_3	n

Et du tableau des *effectifs « théoriques »* :

Valeurs	1	2	3	Total
Effectifs théoriques	$n/3$	$n/3$	$n/3$	n

On désire mesurer l'écart, la *« distance »* entre les deux répartitions. Par analogie avec le calcul d'une distance en géométrie analytique dans l'espace où :

$$d^2 = (x - x')^2 + (y - y')^2 + (z - z')^2$$

1 Membres du groupe : Acker Emmanuelle, Audéoud Jérôme, Athlag Mohamed, Dupuis Claire, Koch Bernard, Quelen Jean-Paul, Weil Dominique.

on calcule ici :

$$D^2_{obs} = (e_1 - \frac{n}{3})^2 + (e_2 - \frac{n}{3})^2 + (e_3 - \frac{n}{3})^2$$

On simule n expériences conformes au modèle équiprobable. On répète un grand nombre de fois cette simulation en calculant à chaque fois D^2_{sim} (en utilisant les effectifs obtenus par simulation). On approche ainsi la loi de probabilité de la variable aléatoire D^2.

Si, lors de ces simulations, la valeur de D^2_{obs} est dépassée très rarement (par exemple dans moins de 5% ou 10% des cas) on rejette l'hypothèse que le modèle équiprobable est adéquat pour décrire l'expérience observée.

Si c'est une distribution de fréquences résultant de n lancers que l'on désire comparer à la loi de probabilité :

Valeurs	1	2	3	Total
Fréquences	f_1	f_2	f_3	1

Valeurs	1	2	3	Total
Probabilités	1/3	1/3	1/3	1

on calculera plutôt :

$$d^2_{obs} = (f_1 - \frac{1}{3})^2 + (f_2 - \frac{1}{3})^2 + (f_3 - \frac{1}{3})^2$$

et on observera par simulations la distribution de la variable aléatoire d^2.

Si, lors de ces simulations, la valeur de d^2_{obs} est dépassée très rarement (par exemple dans moins de 5% ou 10% des cas) on rejette l'hypothèse d'équiprobabilité.

Bien entendu comme $f_i = \frac{e_i}{n}$ on a : $d^2_{obs} = \frac{1}{n^2} D^2_{obs}$ et les deux procédures sont équivalentes.

Voici, dan l'encadré ci-dessous, l'énoncé du problème de Pondichéry.

Nous nous attacherons ensuite à un corrigé détaillé de celui-ci.

L'énoncé de Pondichery

Un pisciculteur possède un bassin qui contient trois variétés de truites : *communes, saumonées* et *arc-en-ciel*. Il voudrait savoir s'il peut considérer que son bassin contient autant de truites de chaque variété. Pour cela il effectue, au hasard, 400 prélèvements d'une truite avec remise et obtient les résultats suivants :

Variétés	Commune	Saumonée	Arc-en-ciel
Effectifs	146	118	136

1. a. Calculer les fréquences de prélèvement f_c d'une truite commune, f_s d'une truite saumonée et f_a d'une truite arc-en-ciel. On donnera les valeurs décimales exactes.

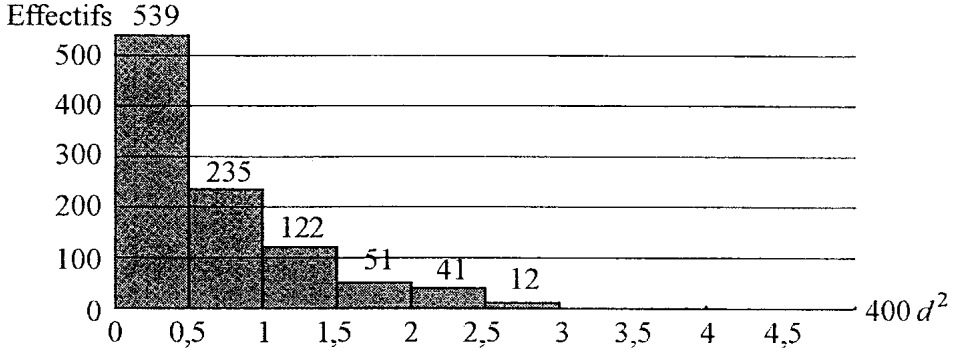
b. On pose :

$$d_{obs}^2 = (f_1 - \frac{1}{3})^2 + (f_2 - \frac{1}{3})^2 + (f_3 - \frac{1}{3})^2$$

Calculer $400d^2$ arrondi à 10^{-2} ; on note $400d_{obs}^2$ cette valeur.

2. A l'aide d'un ordinateur, le pisciculteur simule le prélèvement au hasard de 400 truites suivant la loi équirépartie. Il répète 1000 fois cette opération et calcule à chaque fois la valeur de $400d^2$.

Le diagramme à bandes ci-dessous représente la série des 1000 valeurs de $400d^2$, obtenues par simulation.



Déterminer une valeur approchée à 0,5 près par défaut, du neuvième décile D9 de cette série.

3. En argumentant soigneusement la réponse dire si on peut affirmer avec un risque d'erreur inférieur à 10 % que « le bassin contient autant de truites de chaque variété ».

4. On considère désormais que le bassin contient autant de truites de chaque variété. Quand un client se présente, il prélève au hasard une truite du bassin.

Trois clients prélèvent chacun une truite. Le grand nombre de truites du bassin permet d'assimiler ces prélèvements à des tirages successifs avec remise.

Calculer la probabilité qu'un seul des trois clients prélève une truite commune.

Un corrigé commenté

Question 1 :

Les calculs ne sont pas difficiles mais bien plus commodes si on dispose d'un tableur :

Variétés	Commune	Saumonée	Arc en ciel	Total
Effectifs	146	118	136	400
Fréq.	0,365	0,295	0,340	1
Ecarts	0,03166667	0,03166667	0,03166667	d^2
Carrés	0,00100278	0,00100278	0,00100278	0,00300833
			$400*d^2$	1,20

Les fréquences sont décimales il est donc possible d'en donner une valeur exacte comme le demande l'énoncé.

La multiplication par 400 peut sembler arbitraire. Elle permet ici d'obtenir des valeurs plus lisibles. Ce changement d'échelle n'est pas indispensable si on procède par simulation dans la suite et le graphique donné dans la question 2 pouvait tout aussi bien comporter en abscisse les valeurs de d^2.

On démontre qu'en fait $3\times n\times d^2$ (où n est le nombre de tirages, donc ici 400) suit asymptotiquement une loi du Khi-deux à 2 degrés de liberté. La consultation d'une table du Khi-deux fournit un neuvième décile de 4,6 qui, divisé par 3, donne bien une valeur proche des 1,5 déterminés à la question 2.

Question 2 :

Le cumul, en partant de la fin, des effectifs indiqués donne 12, 53 puis 104 alors que le dixième de l'effectif total est 100. Le neuvième décile est donc dans l'intervalle [1.5 ; 2] et 1.5 en est une valeur approchée à 0,5 près par défaut.

Comme c'est une valeur par défaut, les valeurs simulées dépassent 1,5 dans un peu plus de 10% des cas et si on était amené à rejeter l'hypothèse d'équiprobabilité par la suite (ce ne sera pas le cas), la probabilité de rejet à tort serait elle aussi légèrement supérieure à 0,1.

Question 3 :

La valeur $400d^2_{obs}$ observée est de 1,2. Elle est inférieure (assez nettement) au neuvième décile arrondi qui est 1,5. On ne peut donc rejeter l'hypothèse d'équirépartition.

L'énoncé souhaite une argumentation soigneuse et utilise le terme de « risque d'erreur ». Si on interprète cette expression comme « probabilité de prendre une mauvaise décision » on se heurte à une difficulté majeure qui rend, dans le cas de l'énoncé, toute réponse impossible. Certes les considérations qui suivent n'auront pas, nous l'espérons, trop troublé les candidats.

Il y a en effet deux risques d'erreur :

Lorsque l'on teste une hypothèse (traditionnellement appelée « hypothèse nulle » et notée H0), on procède à des calculs de probabilités en supposant qu'elle est vraie. Les seules probabilités que l'on peut alors calculer sont des probabilités calculées dans le modèle où l'hypothèse H0 est vraie. On peut ainsi calculer la probabilité de *rejeter à tort* l'hypothèse H0 (alors qu'elle est vraie)[2] en adoptant une certaine règle de décision. La

2 appelé risque de première espèce et souvent noté α dans la théorie des tests d'hypothèses.

probabilité d'*accepter à tort* l'hypothèse H0 (alors qu'elle est fausse)[3] ne pourrait se calculer que dans le vrai modèle inconnu dans la pratique.

Dans le cas du test d'équiprobabilité concernant les trois variétés de truites proposé dans l'énoncé, la procédure consiste à prélever 400 truites avec remise, à calculer les fréquences et la quantité $400d^2_{obs}$. La règle de décision, suggérée par l'énoncé et adoptée dans les corrigés, consiste à rejeter l'hypothèse d'équiprobabilité si $400d^2_{obs} > 1{,}5$. Cette dernière valeur étant le neuvième décile (arrondi) des 1000 valeurs de $400d^2$ obtenues par une simulation dans le modèle équiprobable. Le calcul de $400d^2_{obs}$ avec les fréquences observées donne un résultat inférieur à ce D9. On ne peut donc rejeter l'hypothèse d'équiprobabilité.

Si les proportions des truites des trois variétés ne sont pas égales, seule la connaissance des proportions réelles permet de calculer (ou tout du moins d'estimer par simulation) la probabilité d'accepter à tort l'hypothèse d'équiprobabilité en appliquant la règle de décision choisie.

Exemples :

Supposons que les proportions réelles (inconnues dans la pratique) soient :

Communes(1)	Saumonées(2)	Arc-en-ciel(3)
0,31	0,31	0,38

3 appelé risque de deuxième espèce et souvent noté β dans la théorie des tests.

Ce n'est pas l'équirépartition, sans en être trop grossièrement éloigné ! Avec ce modèle, 1000 simulations de 400 tirages avec calcul de $400d^2_{sim}$ donnent la répartition suivante :

0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5	100	Total
140	170	163	124	123	81	52	39	38	24	46		**1000**

On remarque que 473 tirages ont donné $400d^2_{sim}$ inférieur à 1,5. La probabilité que l'on a, avant de faire le test, d'accepter (à tort) l'hypothèse d'équiprobabilité est de 0,473 alors que la probabilité de rejeter (à raison) l'équiprobabilité est 1 – 0,473 soit 0.527.

Plus les proportions réelles sont proches de l'équirépartition, plus la probabilité d'accepter l'hypothèse d'équiprobabilité est élevée.

Supposons que les proportions réelles soient :

Communes(1)	Saumonées(2)	Arc-en-ciel(3)
0,32	0,33	0,35

Avec ce modèle, 1000 simulations de 400 tirages avec calcul de $400d^2_{sim}$ donnent la répartition suivante :

0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5	100	Total
424	229	160	82	46	23	21	5	5	1	4		**1000**

Cette fois ce sont 813 tirages qui ont donné $400d^2_{sim}$ inférieur à 1,5. La probabilité que l'on a, avant de faire le test, d'accepter (à tort) l'hypothèse d'équiprobabilité est de 0,813.

Dans le cas de proportions réelles pratiquement équilibrées comme, par exemple :

Communes(1)	Saumonées(2)	Arc-en-ciel(3)
0,33	0,33	0,34

Une répartition de 1000 valeurs $400d^2_{sim}$ obtenues par simulation est :

0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5	100	Total
508	259	126	64	22	14	4	2	0	1	0		**1000**

La probabilité que l'on a, avant de faire le test, d'accepter (à tort) l'hypothèse d'équiprobabilité est de 0,893. Il est normal que cette valeur soit très proche de 0,9 qui est la probabilité d'accepter l'équiprobabilité lorsqu'elle est vraie.

Dans le cas de proportions réelles plus nettement déséquilibrées comme, par exemple :

Communes(1)	Saumonées(2)	Arc-en-ciel(3)
0,25	0,40	0,35

On observe une répartition des $400d^2_{sim}$ qui permet d'estimer la probabilité d'accepter à tort l'équiprobabilité à 0,040.

0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5	100	Total
2	14	24	24	56	58	73	83	81	78	507		**1000**

Le risque d'erreur (la probabilité de prendre une mauvaise décision à savoir l'acceptation à tort) lorsque la répartition n'est pas uniforme dépend donc de la répartition réelle qui est, par essence, inconnue dans le problème pratique. Plus la répartition réelle est déséquilibrée, plus le test a des chances de le détecter comme l'illustre le tableau récapitulatif des exemples précédents :

Répartition	Probabilité d'acceptation (à tort)
0.33/0.33/0.34	0.893
0.32/0.33/0.35	0.813
0.31/0.31/0.38	0.473
0.25/0.35/0.40	0.040

La seule majoration générale que l'on peut donner dans le cas de cet exemple c'est que la probabilité d'acceptation à tort est toujours inférieure à 0,9.

La prudence incite donc à donner une réponse comme :

« On ne peut rejeter l'hypothèse d'équiprobabilité »
« Rien ne permet de rejeter l'hypothèse d'équiprobabilité »

en laissant dans le flou le « risque d'erreur » dans ce cas. La formulation de la question n'incitait pas les candidats à cette prudence.

Il nous reste à espérer qu'aucun candidat n'aura éprouvé de scrupules à répondre en convertissant sous forme affirmative la question posée et qu'aucun jury n'aura pénalisé l'utilisation d'une forme plus imprécise mais moins sujette à controverse.