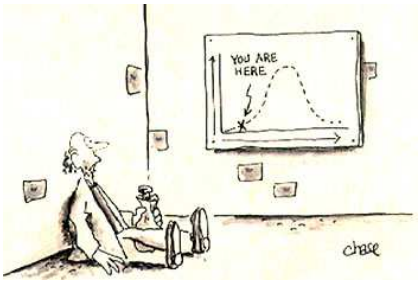
DIX-NEUVIÈME LEÇON

Adéquation à une loi équirépartie



I - De quoi s'agit-il ?

Ce chapitre vous introduit dans le monde des *Statistiques inférentielles* (**Inférer** : tirer une conséquence de quelque proposition, de quelque fait, etc.) auquel vous êtes en fait constamment confronté(e) en tant que citoyen(ne) d'une société hautement médiatisée.
La statistique inférentielle est en effet la science qui permet de « *modéliser une partie observable du réel comme résultant d'un phénomène aléatoire pour lequel on envisage non pas une mais toute une famille de lois de probabilités possibles* » (J.P. Raoult - dossier APMEP statistique inférentielle - Déc 2005)
C'est un sujet très intéressant, exigeant une réflexion approfondie en classe sur les notions abordées, les conclusions à émettre, entraînant un débat intéressant, permettant d'avoir une démarche scientifique partant d'une expérimentation.
Malheureusement, nous sommes loin de disposer du temps nécessaire. Nous nous contenterons donc d'une préparation pure et simple aux exercices d'examen qui sont tous identiques, mis à part le contexte expérimental proposé.

II - Au programme de terminale

On effectue une série d'expériences ayant un nombre fini k d'issues possibles (k modalités).
Par exemple on lance 1 000 fois un dé cubique. Les modalités sont au nombre de six et appartiennent à $\{1;2;3;4;5;6\}$. On veut tester l'**équirépartition** des résultats, c'est à dire qu'on voudrait savoir si on peut ou non rejeter le modéle selon lequel aucune face n'est privilégiée.
En d'autres termes, il sera d'autant plus raisonnable de rejeter l'hypothèse d'équirépartition que la suite des fréquences *observées* $(f_1, f_2, f_3, f_4, f_5, f_6)$ est plus éloignée de la suite constante $\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$.

En effet, nous avons abordé cette année la Loi faible des grands nombres et vous avez observé en seconde que plus le nombre d'expérience est grand, plus les fréquences observées se rapprochent des probabilités. Donc si les fréquences observées pour un nombre important d'expérience s'éloignent d'un modéle « équiréparti » c'est que les probabilités d'obtention de chaque face ne sont sûrement pas égales, et que le modèle équiprobable est à rejeter.

Il est à noter que nous n'obtiendrons ici que des indications nous permettant de rejeter ou non le modèle, mais aucune certitude. Il restera toujours une marge d'erreur que l'on peut quantifier.
Il reste donc à choisir un indicateur d'erreur. On a choisi assez naturellement le carré de la distance euclidienne entre l'expérience et le modèle, c'est à dire qu'on va comparer d défini par

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

à un paramètre de précision arbitrairement choisi (et fourni par les énoncés de Bac).

Attention à la conclusion ! Si d est inférieur (supérieur) au paramètre e, cela veut dire qu'on ne peut pas (peut), avec un risque inférieur à e, rejeter l'hypothèse d'équirépartition. Cela **NE** veut **PAS** dire que l'on peut accepter le modèle.

III - Un exemple commenté

Reprenons notre dé cubique. On a lancé un dé 200 fois et on a obtenu les résultats suivants :

Face	1	2	3	4	5	6
Nombre de sorties de chaque face	28	33	34	29	26	50

On calcule alors le carré de la distance associée, qu'on notera d_{obs}

$$d_{obs}^2 = \left(\frac{28}{200} - \frac{1}{6}\right)^2 + \left(\frac{33}{200} - \frac{1}{6}\right)^2 + \left(\frac{34}{200} - \frac{1}{6}\right)^2 + \left(\frac{29}{200} - \frac{1}{6}\right)^2 + \left(\frac{26}{200} - \frac{1}{6}\right)^2 + \left(\frac{50}{200} - \frac{1}{6}\right)^2 \approx 0,00948$$

Que faire alors de ce résultat ?
Il nous faudrait alors une table de distances sur un plus grand nombre d'expériences pour avoir une échelle de comparaison. On effectue donc 1 000 simulations de 200 lancers de dé (sur XCAS on entre `rand(7)` pour avoir un nombre entier entre 1 et 6). On obtient alors 1 000 valeurs de d^2.Ces simulations ne peuvent évidemment pas être faites le jour du Bac. On vous fournit donc les résultats, enfin une partie.
Par exemple, on peut vous dire que le $99^{ème}$ centile vaut 0,01258 : qu'est ce que ça veut dire ?
Eh bien cela signifie que 99% des valeurs obtenues pour d^2 sont inférieures à 0,01258.
Notons d_{99}^2 cette valeur. On a donc

$$d_{obs}^2 < d_{99}^2$$

On en déduit qu'on ne peut pas, au seuil de 1%, rejeter l'hypothèse d'équirépartition.

Maintenant, il est précisé que le $95^{ème}$ centile vaut 0,00888. Cette fois

$$d_{obs}^2 > d_{95}^2$$

Cette fois-ci on peut, au risque d'erreur de 1%, rejeter l'hypothèse de régularité du dé.

Les valeurs des déciles pourront être données directement ou lues sur un histogramme ou une boîte à moustache (ça s'est vu...).

IV - Exercices

Exercice 1 Bac : π

Les 1 000 premières décimales de π sont données ici par un ordinateur :
1415926535 8979323846 2643383279 5028841971 6939937510 5820974944 5923078164 0628620899 8628034825 3421170679
8214808651 3233066470 9384460959 0582235725 3594085234 8111745028 4102701930 5211055596 4462294895 4930301964
4288109756 6593344612 8475648233 7867831652 7120190914 5648566923 4603486534 5432664825 3393607260 2491412737
2450700660 6315580574 8815209209 6282925409 1715364367 8925903600 1133053054 8820466525 3841469519 4151160943
3057270365 7595919530 9218611738 1932611793 1051185480 7446297996 2749567355 8857527240 9122793318 3011949129
8336733624 4065664308 6025394946 3952247371 9070217986 0943702770 5392171762 9317675238 4674818467 6691051320
0056812714 5263560827 7857753427 9778900917 3637178721 4684409012 2495343054 6549585371 0507922796 8925892354
2019956112 1290219608 6403441815 9813629774 7713099605 1870721134 9999998372 9780499510 5973173281 6096318599
0244594553 4690830264 2522300253 3446850352 6193110017 1010003137 8387528865 8753320830 1420617177 6691473035
9825349042 8755460731 1595620633 8235378759 3751957781 8577805321 7122600661 3001927876 6111959092 1642019894
En groupant par valeurs entre 0 et 9 ces décimales, on obtient le tableau suivant :

Valeurs	0	1	2	3	4	5	6	7	8	9
Occurrences	93	116	102	102	94	97	94	95	101	106

Avec un tableur, on a simulé 1 000 expériences de 1 000 tirages aléatoires d'un chiffre compris entre 0 et 9.

Pour chaque expérience, on a calculé $d^2 = \sum_{k=0}^{k=9} \left(f_k - 0,1\right)^2$ où f_k représente, pour l'expérience, la fréquence observée du chiffre k.
On a alors obtenu une série statistique pour laquelle on a calculé le premier et neuvième décile (d_1 et d_9), le premier et troisième quartile (Q_1 et Q_3) et la médiane (Me) :
$d_1 = 0,000\,422$; $Q_1 = 0,000\,582$; Me $= 0,000\,822$; $Q_3 = 0,001\,136$; $d_9 = 0,001\,45$.

En effectuant le calcul de d_2 sur la série des 1 000 premières décimales de π, on obtient :

☐ $0,000\,456$ ☐ $0,004\,56$ ☐ $0,000\,314$

Un statisticien découvrant le tableau et ignorant qu'il s'agit des décimales de π, fait l'hypothèse que la série est issue de tirages aléatoires indépendants suivant une loi équirépartie. Il prend un risque de 10 % de rejeter cette hypothèse quand elle est vraie. Accepte-t-il cette hypothèse ?

☐ Oui ☐ Non ☐ Il ne peut pas conclure.

Solution 1 Bac : π
On calcule :
$d^2 = (0,093-0,1)^2 + (0,116-0,1)^2 + (0,102-0,1)^2 + (0,102-0,1)^2 + (0,094-0,1)^2 + (0,094-0,1)^2 + 0,097-0,1)^2 + (0,095-0,1)^2 + 0,101-0,1)^2 + (0,106-0,1)^2 = 0,000\,456$.
On a $0,000\,456 < 0,001\,45$ soit $d^2 < d_9$, donc il accepte cette hypothède avec un risque de 10 % de la rejeter.

Exercice 2 Bac : Dé tétraèdrique

Partie A
On dispose d'un dé en forme de tétraèdre régulier, possédant une face bleue, deux faces rouges et une face verte ; on suppose le dé parfaitement équilibré.
Une partie consiste à effectuer deux lancers sucessifs et indépendants de ce dé. À chaque lancer on note la couleur de la face cachée.
On considère les évènements suivants :
E est l'évènement « à l'issue d'une partie, les deux faces notées sont vertes »,
F est l'évènement « à l'issue d une partie, les deux faces notées sont de la même couleur ».

1. Calculer les probabilités des évènements E et F ainsi que la probabilité de E sachant F.
2. On effectue dix parties identiques et indépendantes.
   Calculer la probabilité d'obtenir au moins deux fois l'évènement F au cours de ces dix parties (on en donnera une valeur approchée décimale à 10^{-3} près).

Partie B
On souhaite savoir si le dé utilisé peut être considéré comme parfaitement équilibré.
Pour cela on numérote de 1 à 4 les quatre faces de ce dé, puis on lance, ce dé 160 fois en notant le nombre n_i de fois où chaque face est cachée ; on obtient les résultats suivants :

face i	1	2	3	4
effectif n_i	34	48	46	32

On note f_i la fréquence relative à la face n_i et d^2_{obs} le réel $\sum_{i=1}^{4} \left(f_i - \frac{1}{4}\right)^2$.
On simule ensuite 1 000 fois l'expérience consistant à tirer un chiffre au hasard 160 fois parmi l'ensemble (1 ; 2 ; 3 ; 4) puis, pour chaque simulation, on calcule
$d^2 = \sum_{i=1}^{4} \left(F_i - \frac{1}{4}\right)^2$, où F_i est la fréquence d'apparition du nombre i. Le 9$^{\text{e}}$ décile de la série statistique des 1 000 valeurs de d^2 est égal à 0,009 8.
Au vu de l'expérience réalisée et au risque de 10 %, peut-on considérer le dé comme parfaitement équilibré ?

Exercice 3 Bac : boîte à moustache

Les guichets d'une agence bancaire d'une petite ville sont ouverts au public cinq jours par semaine : les mardi, mercredi, jeudi, vendredi et samedi.
Le tableau ci-dessous donne la répartition journalière des 250 retraits d'argent liquide effectués aux guichets une certaine semaine.

Jour de la semaine	mardi	mercredi	jeudi	vendredi	samedi
Rang i du jour	1	2	3	4	5
Nombre de retraits	37	55	45	53	60

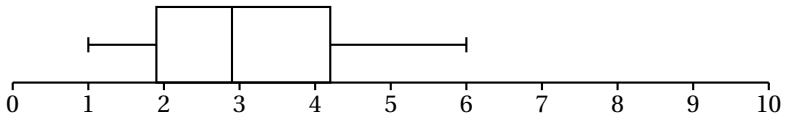
On veut tester l'hypothèse « le nombre de retraits est indépendant du jour de la semaine ». On suppose donc que le nombre des retraits journaliers est égal à $\frac{1}{5}$ du nombre des retraits de la semaine.

On pose $d^2_{\text{obs}} = \sum_{i=1}^{5} \left(f_i - \frac{1}{5} \right)^2$ où f_i est la fréquence des retraits du i-ème jour.

1. Calculer les fréquences des retraits pour chacun des cinq jours de la semaine.

2. Calculer alors la valeur de $1\,000 d^2_{\text{obs}}$ (la multiplication par 1 000 permet d'obtenir un résultat plus lisible).

3. En supposant qu'il y a équiprobabilité des retraits journaliers, on a simulé 2 000 séries de 250 retraits hebdomadaires.
Pour chaque série, on a calculé la valeur du $1\,000 d^2_{\text{obs}}$ correspondant. On a obtenu ainsi 2 000 valeurs de $1\,000 d^2_{\text{obs}}$.
Ces valeurs ont permis de construire le diagramme en boîte ci-dessous où les extrémités des « pattes » correspondent respectivement au premier décile et au neuvième décile.



Lire sur le diagramme une valeur approchée du neuvième décile.

4. En argumentant soigneusement la réponse, dire si pour la série observée au début, on peut affirmer, avec un risque d'erreur inférieur à 10 %, que « le nombre de retraits est indépendant du jour de la semaine » ?

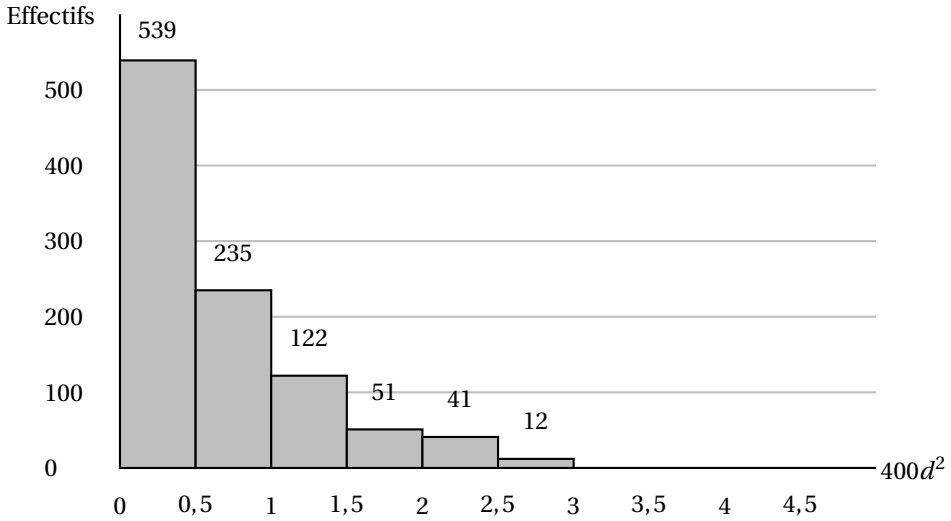
Exercice 4 Bac : histograme

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 :

Variété	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 $d^2 = \left(f_c - \frac{1}{3} \right)^2 + \left(f_s - \frac{1}{3} \right)^2 + \left(f_a - \frac{1}{3} \right)^2$.
Calculer $400d^2$ arrondi à 10^{-2} ; on note $400 d^2_{\text{obs}}$ cette valeur.
À l'aide d'un ordinateur, le pisciculteur simule le prélèvement au hasard de 400 truites suivant la loi équirépartie. Il répète 1 000 fois cette opération et calcule à chaque fois la valeur de $400d^2$.
Le diagramme à bandes ci-dessous représente la série des 1 000 valeurs de $400d^2$, obtenues par simulation.



2. 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.