Leçon 13

Adéquation *à* une loi équirépartie

I) Principe du test sur un exemple détaillé

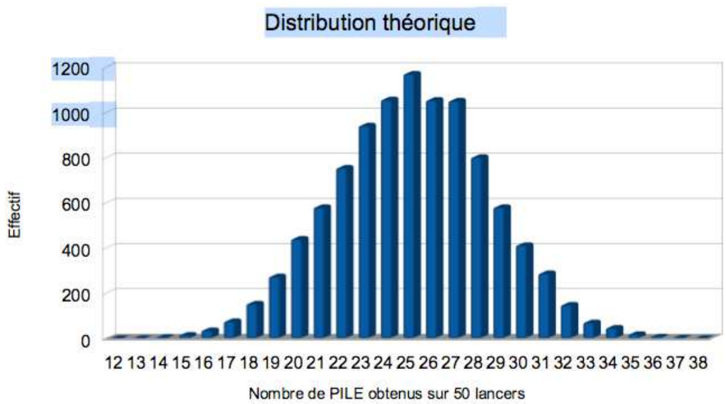
Cette partie vous introduit dans le monde des *Statistiques inférentielles* (**Inférer** : tirer une conséquence de quelque proposition, de quelque fait, etc.) auquel vous êtes en fait constamment confronté(e) en tant que citoyen(ne) d'une société hautement médiatisée. La statistique inférentielle est en effet la science qui permet de « *modéliser une partie observable du réel comme résultant d'un phénoméne aléatoire pour lequel on envisage non pas une mais toute une famille de lois de probabilités possibles* » (J.P. Raoult - dossier APMEP- Déc 2005)
C'est un sujet très intéressant, exigeant une réflexion approfondie 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.
Dans tous les cas, il s'agira d'étudier les résultats d'une série d'expériences ayant un nombre fini k d'issues possibles.
Regardons en détail un **exemple**.

On dispose d'une pièce de monnaie et on souhaite savoir si elle est bien équilibrée. Mais comment le savoir ?

On la lance 50 fois (à la main, plus devient fastidieux). On a obtenu 30 fois Pile et 20 fois Face.

Peut-on raisonnablement estimer que la pièce est équilibrée ?

En fait, on ne peut jamais être sûr de rien, mais d'un point de vue statistique, on peut avoir une idée de la réponse, avec une certaine marge d'erreur, grâce aux simulations.
En effet, on peut simuler par un ordinateur 10 000 fois l'expérience de 50 lancers, pour une pièce équilibrée (puisqu'on veut tester l'hypothèse d'équirépartition de notre propre pièce). On l'a fait sur tableur et on a obtenu :



Nombre de "PILE" obtenus	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Effectif	1	1	6	16	36	76	154	274	441	581	755	942	1057	1171

Nombre de "PILE" obtenus	26	27	28	29	30	31	32	33	34	35	36	37	38
Effectif	1055	1052	802	581	413	288	149	71	47	19	8	3	1

On constate que sur cette simulation, à peine plus d'un dixième des expériences donnent exactement 25 Piles et 25 Faces. De plus, pour beaucoup d'entre elles, le nombre de Pile et de Face est assez éloigné de cette répartition, alors que la pièce est équilibrée. On parle de **fluctuation d'échantillonnage**.
On constate également qu'une certaine proportion (pas forcément négligeable) de ces 10 000 expériences a obtenu 30 Pile et 20 Face. Cet événement obtenu lors de notre propre expérience n'est donc pas si improbable que cela pour une pièce équilibrée.
Mais on aimerait décider si notre pièce est, elle aussi, équilibrée, au risque de commettre une erreur.

Quels peuvent-être les critères de décisions ? Quelles sont alors les erreurs commises ?

Il est clair que plus les fréquences observées $\left(f_{Pile}, f_{\text{Face}}\right)$ sont éloignées des fréquences théoriques $\left(\frac{1}{2}, \frac{1}{2}\right)$, plus il est raisonnable de rejeter l'hypothèse d'équirépartition. Mais cela ne suffit pas pour décider.
On utilise comme indicateur d'erreur le carré de la distance euclidienne entre l'expérience et le modèle. En fait, on va comparer d^2_{obs} défini par

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

à un paramètre de précision e arbitrairement choisi (et fourni par les énoncés de Bac), et dépendant de la simulation.
Ici on obtient

$$d^2_{\text{obs}} = \left(f_{\text{P}ile} - p(\text{P}ile)\right)^2 + \left(f_{\text{F}ace} - p(\text{F}ace)\right)^2 = \left(\frac{30}{50} - \frac{1}{2}\right)^2 + \left(\frac{20}{50} - \frac{1}{2}\right)^2 = \frac{1}{50} = 0.02$$

Evidemment, plus une épreuve est proche des fréquences théoriques, plus son d^2_{obs} est proche de 0 et inversement.

Avec le tableur, on calcule d^2 pour chacune des simulations de 50 lancers, et on fait le tableau des effectifs cumulés croissants de ces d^2. (Notons que le d^2 pour 26 piles est le même que celui pour 24 piles), et ainsi de suite.

Nombres de piles	25/25	24/26	23/27	22/28	21/29	20/30	19/31	18/32	17/33	16/34	15/35	14/36	13/37	12/38
Valeurs d^2	0	$\frac{1}{1250}$	$\frac{2}{625}$	$\frac{9}{1250}$	$\frac{8}{625}$	$\frac{1}{50}$	$\frac{18}{625}$	$\frac{49}{1250}$	$\frac{32}{625}$	$\frac{81}{1250}$	$\frac{2}{25}$	$\frac{121}{1250}$	$\frac{72}{625}$	$\frac{169}{1250}$
Effectif cumulés croissants	1171	3283	5277	6834	7996	8850	9412	9715	9862	9945	9980	9994	9998	10000

Maintenant, nous allons décider suivant une marge d'erreur fixée par l'énoncé, si nous considérons que notre pièce est équilibrée ou non. Pour cela, tout dépend en fait de la position de notre d^2_{obs} dans le tableau ci-dessus.
Par exemple, si on se donne une marge d'erreur de 10%, cela signifie que l'on considère que les 10% de d^2 les plus élévés obtenus par notre simulation sont marginaux.
Ainsi, si notre d^2_{obs} obtenu par notre expérience en fait partie, on rejettera notre hypothèse d'équirépartition, sinon, on ne pourra pas la rejeter (et par abus de langage, il est parfois dit qu'on la valide ou qu'on l'accepte).

Pour une marge d'erreur de 10%, on raisonne par rapport au 9^e décile. En effet, le 9^e décile D_9 est défini comme le réel tel qu'au moins 90% des d^2 soient inférieurs ou égaux à D_9. En général, l'énoncé fourni directement les quartiles, déciles, centiles ... utiles pour répondre. Mais on peut aussi nous donner le tableau ci-dessus, construire la courbe des effectifs cumulés croissants et le retrouver graphiquement (improbable), ou encore, l'énoncé peut fournir un diagramme à moustache.

Règle de décision

- Si $d^2_{\text{obs}} > \text{D}_9$, alors on rejette l'hypothèse d'équirépartition pour notre pièce, avec une marge d'erreur de 10%.
- Si $d^2_{\text{obs}} < \text{D}_9$, alors on ne peut pas rejeter l'hypothèse d'équirépartition pour notre pièce, avec une marge d'erreur de 10%.

Ici, $\text{D}_9 = 0.022$. Cela signifie donc qu'au moins $\frac{9}{10}$ des valeurs de d^2 sont inférieures ou égales à $\text{D}_9 = 0.022$.
Comme $d^2_{\text{obs}} < \text{D}_9$ donc on ne peut pas rejeter l'hypothèse d'équirépartition, avec une marge d'erreur de 10%.
Si on nous avait donné le 95^e centile, on aurait effectué un test avec une marge d'erreur de 5%, etc.

Et si nous avions obtenu 32 Pile et 18 Face avec une marge d'erreur de 10% ?

Même question avec une marge d'erreur de 1% (On nous donne le 99^e centile : $C_{99} = \frac{3}{5}$)

Remarque : Ceci paraît paradoxal, mais en fait, une marge d'erreur de 1% signifie que l'on considère que seulement 1% des résultats de la simulation sont marginaux. Il est donc moins probable que notre observation tombe dans ces 1% que dans 10%. La marge d'erreur à 1% a donc tendance à valider « trop » d'observations comme équirépartie.
Et le cas marge d'erreur à 0% ?
A l'inverse, un modèle avec une marge d'erreur de 90%, on rejetterait « trop » d'observations.
Certains statisticiens considèrent douteux et trop beaux pour être vrai de tels modèles.

II) Exemples

On a lancé un dé 200 fois un dé cubique 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

1000 simulations de 200 lancers de dé équilibré ont été réalisées sur un tableur. Pour chaque expérience, on a calculé $d^2 = \sum_{k=1}^{6} \left(f_k - \frac{1}{6}\right)^2$ où f_k représente pour l'expérience, la fréquence observée du numéro obtenu k. On a alors obtenu une série statistique pour laquelle on a calculé le 99^e centile qui vaut 0,01258 et le 95^e centile qui vaut 0,00888.
Peut-on accepter l'hypothèse d'un dé équilibré au seuil de 1% ? 5% ?

Exercice 1. Bac : Dé tétraédrique

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. A 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^e décile de la série statistique des 1 000 valeurs de d^2 est égal à 0,0098.
Au vu de l'expérience réalisée et au risque de 10 %, peut-on considérer le dé comme parfaitement équilibré ?

Exercice 2. Bac : boîte à moustache

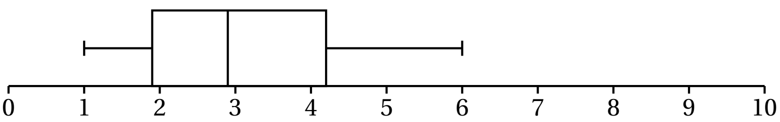
Les guichets d'une agence bancaire d'une petite ville sont ouverts au public cinq jours par semaine : les mardi, mercredi, jeudi, vendredi et samedi.
Le tableau ci-dessous donne la répartition journalière des 250 retraits d'argent liquide effectués aux guichets une certaine semaine.

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 $\dfrac{1}{5}$ du nombre des retraits de la semaine.

On pose $d^2_{\text{obs}} = \displaystyle\sum_{i=1}^{5}\left(f_i - \frac{1}{5}\right)^2$ 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 $1\,000d^2_{\text{obs}}$ (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 $1\,000d^2_{\text{obs}}$ correspondant. On a obtenu ainsi 2 000 valeurs de $1\,000d^2_{\text{obs}}$.
   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 3. (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 8111745 …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 $d^2 = \displaystyle\sum_{k=0}^{k=9}\left(f_k - 0,1\right)^2$ 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.