Chapitre 10

Adéquation à une loi équirépartie

Sommaire

10.1 Le contexte

D'après bac à maths.

Imaginons que l'on dispose d'une pièce de monnaie dont on souhaite savoir si elle est bien équilibrée ou non.

On la lance 50 fois (on peut difficilement envisager de la lancer un très grand nombre de fois à la main) et on obtient, par exemple, 30 fois PILE et 20 fois FACE. Peut-on raisonnablement estimer que la pièce est bien équilibrée ou non ?

En fait, on ne peut être sûr de rien mais on peut, d'un point de vue statistique, avoir une idée plus ou moins valide de la réponse grâce aux simulations. En effet, on peut simuler (avec l'ordinateur par exemple) 10 000 fois l'expérience consistant à lancer 50 fois une pièce de monnaie bien équilibrée, et examiner si obtenir 30 fois PILE et 20 fois FACE est exceptionnel ou non, et dans quelle mesure. On pourra ainsi décider si notre piece est, elle aussi, bien équilibrée (au risque de commettre une erreur).

Nous sommes donc confrontés à plusieurs questions :

quels sont les critères de décisions ?
quel est le risque d'erreur ?

Voilà pour le principe.

Voyons maintenant, plus en détails, les calculs et la méthode.

Notons $\mathscr{E}$ l'expérience qui consiste à lancer 50 fois une pièce de monnaie. On simule, sur ordinateur, 10 000 fois l'expérience $\mathscr{E}$ pour une pièce bien équilibrée (hypothèse d'équirépartition : chaque face à une probabilité de $\frac{1}{2}$ d'apparition) et on obtient les résultats donnés dans le tableau 10.1 page suivante et illustrés dans le diagramme en dessous.

Par exemple, sur les 10 000 expériences, 76 ont donné 17 PILE (et donc 33 FACE). On constate que 413 ont donné 30 PILE (et donc 20 FACE) ce qui répresente tout de même 4,13 % de l'effectif total. On constate, sur cette simulation, qu'à peine plus de un dixième de l'effectif donne exactement 25 PILE et 25 FACE et que de nombreuses expériences montrent que sur 50 lancers, le nombre de PILE ou FACE n'est pas forcément voisin, bien que les calculs aient été faits pour une pièce bien équilibrée (phénomène de fluctuation d'échantillonnage).

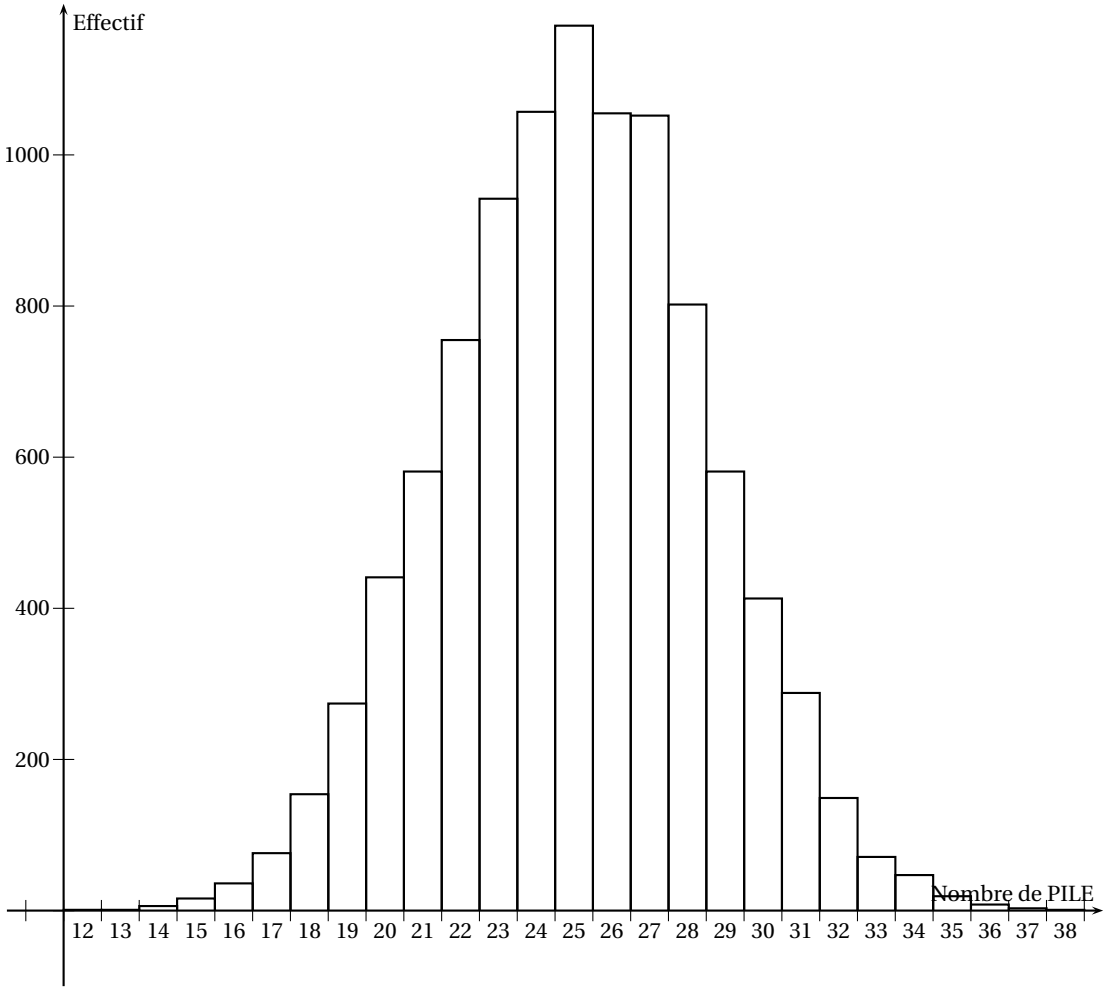
Finalement nos 30 PILE et 20 FACE ne sont peut-être pas si loin que ça de la théorie d'équiprobabilité.

Pour mesurer la *distance* de l'événement « Obtenir 30 PILE et 20 FACE » à la théorie dans le cas d'une pièce équilibrée, nous passons par les fréquences : nous avons obtenu une fréquence de PILE de $\frac{3}{5}$ au lieu de $\frac{1}{2}$, et une fréquence de FACE de $\frac{2}{5}$ au lieu de $\frac{1}{2}$.

TABLE 10.1 – Distribution obtenue suite à la simulation sur ordinateur

Nombre de « PILE » obtenus	12	13	14	15	16	17	18	19	20	21	22	23	24	25
effectifs	1	1	6	16	36	76	154	274	441	581	755	942	1 057	1 171
Nombre de « PILE » obtenus	26	27	28	29	30	31	32	33	34	35	36	37	38	
effectifs	1 055	1 052	802	581	413	288	149	71	47	19	8	3	1	

TABLE 10.2 – Diagramme de la distribution



Nous disposons de deux moyens [1] de mesurer la distance entre deux nombres x et y :

- soit $|x-y|$;
- soit $(x-y)^2$.

Pour des raisons pratiques et de cohérence avec la variance, c'est la seconde distance qui a été choisie par les statisticiens. On la note d^2_{obs} (le carré pour rappeler qu'il s'agit de la distance de l'événement observé à la simulation donnée par le carré de la différence).

On a ainsi :

$$d^2_{\text{obs}} = \left(\frac{3}{5} - \frac{1}{2}\right)^2 + \left(\frac{2}{5} - \frac{1}{2}\right)^2 = 0,02$$

Ce nombre étant souvent très petit, pour des raisons de lisibilité (il est plus facile de comparer, par exemple, 210 et 196 que 0,0210 et 0,0196) on le multiplie parfois par un coefficient arbitraire. Ici nous utiliserons $5\,000 d^2_{\text{obs}}$. Pour l'événement « Obtenir 30 PILE et 20 FACE », $5\,000 d^2_{\text{obs}} = 100$

Calculons la quantité $5\,000 d^2$ pour chaque résultat de la simulation pour savoir si une telle distance est rare. Cela donne :

1. Voir l'introduction des mesures de dispersion en Première.

Nombre de « PILE » obtenus	12	13	14	15	16	17	18	19	20	21	22	23	24	25
$5000d^2$	676	576	484	400	324	256	196	144	100	64	36	16	4	0
effectifs	1	1	6	16	36	76	154	274	441	581	755	942	1 057	1 171
Nombre de « PILE » obtenus	26	27	28	29	30	31	32	33	34	35	36	37	38	
$5000d^2$	4	16	36	64	100	144	196	256	324	400	484	576	676	
effectifs	1 055	1 052	802	581	413	288	149	71	47	19	8	3	1	

Évidemment, plus une épreuve de la simulation est proche des fréquences théoriques, plus sa valeur de $5000d^2$ est proche de 0 et inversement. Ainsi si une épreuve a donné très exactement 25 PILE (et donc 25 FACE), $5000d^2$, c'est-à-dire la distance par rapport à la théorie de l'équiprobabilité, sera de 0 et si une épreuve a donné 50 PILE (et donc 0 FACE), $5000d^2$, c'est-à-dire la distance par rapport à la théorie de l'équiprobabilité, sera de 1 250, c'est-à-dire très grande.

Réorganisons les données en faisant un tableau des valeurs $5000d^2$ suivants les effectifs cumulés croissants :

Valeurs de $5000d^2$	0	4	16	36	64	100	144	196	256	324	400	484	576	676
effectifs cumulés croissants	1 171	3 283	5 277	6 834	7 996	8 850	9 412	9 715	9 862	9 945	9 980	9 994	9 998	10 000

Maintenant, nous allons décider, suivant une marge d'erreur fixée à l'avance, si nous considérons que notre pièce peut être envisagée comme bien équilibrée ou non. Pour cela, tout dépend de la position de notre $5000d^2_{\text{obs}}$ dans le tableau ci-dessus.

La règle de décision usuelle est la suivante :

- si on se donne une marge d'erreur de 10 %, on raisonne par rapport au neuvième décile, D_9, qui est défini comme un réel tel qu'au moins 90 % de l'effectif total ait une valeur inférieure ou égale à D_9 :
  - si $5000d^2_{\text{obs}} \leqslant D_9$, alors on considere que le modèle observe suit la loi d'équirépartition
  - $5000d^2_{\text{obs}} > D_9$, alors on considère que le modèle observé ne suit pas la loi d'équirépartition

  Cela revient à considérer que 90 % des distances à la théorie, celles les plus proches de 0, comme pouvant relever des variations dues à la fluctuation d'échantillonage et à considérer que 10 % des distances à la théorie, celles les plus éloignées de 0, comme pouvant revéler un problème quant à l'adéquation entre le phénomène observé et la loi équirépartie.
- si on se donne une autre marge d'erreur, par exemple 1 %, on raisonne de même en comparant $5000d^2_{\text{obs}}$ au quatre-vingt-dix-neuvième centile.

Dans notre situation, calculons le neuvième décile D_9. Parmi les valeurs de $5000d^2$, on prend celles qui sont inférieures ou égales à la $9000^{\text{ième}}$ distance (il y a 10 000 simulations). D'après le tableau, D_9 est compris entre 100 et 144.
Ici $5000d^2_{\text{obs}} = 100 \leqslant D_9$, donc, nous pouvons affirmer, avec une marge d'erreur de 10 %, que notre pièce est bien équilibrée. Par contre, si nous avions eu une autre pièce donnant 32 PILE et 18 FACE, nous aurions eu dans ce cas :

$$5000d^2_{\text{obs}} = \left(\frac{32}{50} - \frac{1}{2}\right)^2 + \left(\frac{18}{50} - \frac{1}{2}\right)^2 = 196$$

Cette valeur de $5000d^2_{\text{obs}}$ est trop marginale : $5000d^2_{\text{obs}} > D_9$. Une telle pièce serait considérée comme non équilibrée (la distance à la théorie de l'équiprobabilité est trop grande, c'est qu'il ne doit pas s'agir d'une pièce équilibrée) avec une marge d'erreur de 10 %.

Remarques.
- Les résultats peuvent différer si on recommence une autre simulation car l'étendue peut être différente, les valeurs des déciles (ou centiles ou autres) également.
- Il peut arriver qu'on rejette le modèle observé s'il est inférieur au premier décile. Dans ce cas, les fréquences observées sont très proches des fréquences théoriques (puisque d^2 est petit). La probabilité que le modèle observé soit équiréparti est donc très forte. Ceci dit, certains statisticiens le considèrent comme « douteux » (trop beau pour être vrai), c'est ce qui s'est passé pour certains résultats du biologiste Mendel qui avait des résultats statistiques tellement conformes aux fréquences théoriques que certains le soupçonnent aujourd'hui d'avoir embelli ses mesures !

10.2 Bilan

Soit $\mathscr{E}$ l'expérience qui consiste à répéter n fois une épreuve comportant k issues.
On cherche à savoir, si d'après les résultats observés, on peut décider si l'épreuve suit le modèle d'équirépartition.
On note :

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

On suppose que l'on dispose de données simulées (un grand nombre de fois) sur un modèle théoriquement équireparti et on étudie la série statistique des grandeurs d^2 obtenues.
Pour une marge d'erreur de 10 %, on raisonne avec le neuvième decile D_9 de la série des d^2 obtenue par la simulation : si $d^2_{\text{obs}} \leqslant D_9$, on considère que l'expérience observée est équirépartie avec une marge d'erreur de 10 % ; dans le cas contraire, on considère que l'expérience observée n'est pas équirépartie.

10.3 Exercices

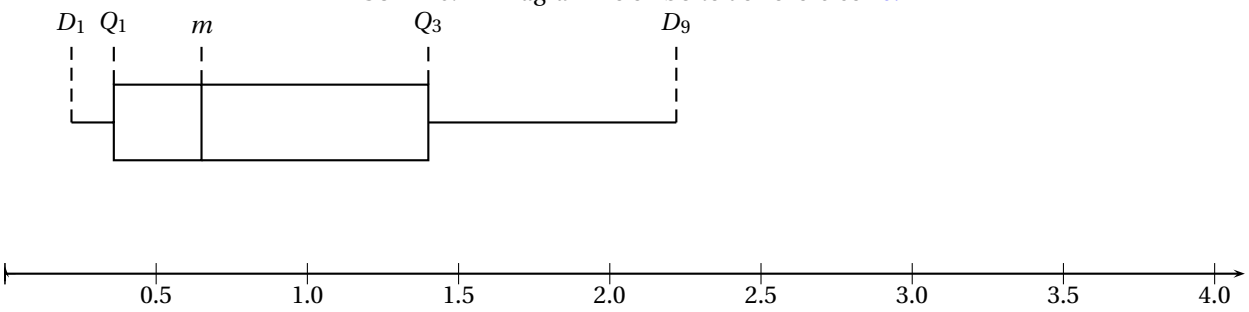
EXERCICE 10.1.
Une clinique fait des statistiques sur les naissances (naturelles et non provoquées) selon le jour de la semaine.
Sur 1 000 naissances naturelles relevées, on obtient les résultats suivants :

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Nombre de naissances	146	163	158	156	156	116	105

On s'intéresse à la validité de l'hypothèse « le nombre de naissance est indépendant du jour de la semaine ». Pour tout entier i compris entre 1 et 7, on note f_i la fréquence des naissances le $i^{\text{ème}}$ jour de la semaine.

1. Calculer $d^2_{\text{obs}} = \sum_{i=1}^{7} \left(f_i - \frac{1}{7} \right)$ puis donner la valeur de $1\,000 d^2_{\text{obs}}$ arrondie à 10^{-2}.
2. On simule sur un ordinateur 50 000 séries de 1 000 naissances équiréparties sur les sept jours de la semaine. Pour chacune de ces 5 000 séries, l'ordinateur a calculé la valeur de $1\,000 d^2$ (où d est la distance entre les fréquences de la série et les fréquences théoriques). Ces valeurs ont permis de construire le diagramme en boîte de la figure 10.1 de la présente page. Avec un risque d'erreur de 10 %, peut-on considérer que le nombre de naissances observées dans la clinique est indépendant du jour de la semaine ?

FIGURE 10.1 – Diagramme en boite de l'exercice 10.1



EXERCICE 10.2.
D'après une étude de l'INSERM, les infarctus ne se produisent pas également chaque jour de la semaine.
Les fréquences calculées sur 17 000 hommes âgés de 25 à 54 ans décédés entre 1 991 et 2 001 sont les suivantes :

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Fréquence	14,8 %	13,3 %	13,4 %	13,6 %	13,8 %	15 %	16,1 %

Reprendre la simulation de l'exercice 10.1 pour confirmer ou infirmer l'étude de l'INSERM.

EXERCICE 10.3.
On dispose d'un dé tétraèdrique et on voudrait savoir s'il est bien équilibré. Pour cela on lance 200 fois le dé et on obtient le tableau suivant :

Face k	1	2	3	4
Nombre de sorties de la face k	58	49	52	41

1. Calculer les fréquences de sorties f_k observées pour chacune des faces.
2. On pose $d^2 = \sum_{k=1}^{4} \left(f_k - \frac{1}{4}\right)^2$. Calculer d^2.
3. On effectue maintenant 1 000 simulations des 200 lancers d'un dé tétraèdrique bien équilibré et on calcule pour chaque simulation le nombre d^2. On obtient pour la série statistique des 1 000 valeurs de d^2 les résultats suivants :

Mimimum	D_1	Q_1	Médiane	Q_3	D_9	Maximum
0,001 24	0,001 92	0,002 35	0,002 81	0,003 45	0,004 52	0,010 15

   Au risque de 10 %, peut-on considérer que ce dé est pipé ?
4. Mêmes questions avec un autre dé ayant donné les résultats suivants :

Face k	1	2	3	4
Nombre de sorties de la face k	30	52	46	32

EXERCICE 10.4.
Un professeur a demandé à ses élèves de jeter 200 fois un dé et de noter leurs résultats.
Voici les relevés de Mathieu :

Face k	1	2	3	4	5	6
Nombre de sorties de la face k	37	34	33	33	29	34

Les résultats lui semblent très probants et Mathieu s'attend à des félicitations.

1. Calculer la valeur de $6\,000 d^2_{\text{obs}}$ dans l'hypothèse d'un dé bien équilibré.
2. Une simulation à l'ordinateur a donné les résultats suivants :

i	5	10	20	30	40	50	100	150
Nombre de simulations où $6\,000 d^2 < i$	36	155	451	698	847	920	995	1000

   Pourquoi le professeur peut-il légitimement penser que Mathieu n'a pas lancé 200 fois le dé ?

EXERCICE 10.5 (Amérique du nord – Juin 2 009)**.**
Un pépiniériste a planté trois variétés de fleurs dans une prairie de quelques hectares : des violettes, des primevères et des marguerites. Il se demande s'il peut considérer que sa prairie contient autant de fleurs de chaque variété. Il cueille au hasard 500 fleurs et obtient les résultats suivants :

Variétés	Violettes	Primevères	Marguerites
Effectifs	179	133	188

1. Calculer les fréquences f_V d'une fleur de variété Violette, f_P d'une fleur de variété Primevère et f_M d'une fleur de variété Marguerite. On donnera les valeurs décimales exactes.
2. On note $d^2_{\text{obs}} = \left(f_V - \frac{1}{3}\right)^2 + \left(f_P - \frac{1}{3}\right)^2 + \left(f_M - \frac{1}{3}\right)^2$.

   Calculer $500 d^2_{\text{obs}}$. On donnera une valeur approchée arrondie au millième.
3. Le pépiniériste, ne voulant pas compter les quelques milliards de fleurs de sa prairie, opère sur ordinateur en simulant le comptage, au hasard, de 500 fleurs suivant la loi équirépartie. Il répète 2 000 fois l'opération et calcule à chaque fois la valeur de $500 d^2_{\text{obs}}$. Ses résultats sont regroupés dans le tableau suivant :

Intervalle auquel appartient $500 d^2_{\text{obs}}$	[0 ; 0,5[	[0,5 ; 1[	[1 ; 1,5[	[1,5 ; 2[	[2 ; 2,5[	[2,5 ; 3[	[3 ; 3,5[	[3,5 ; 4[	[4 ; 4,5[	[4,5 ; 5[
Nombre par intervalle	163	439	458	350	231	161	80	47	37	34

   Par exemple : le nombre $500 d^2_{\text{obs}}$ apparaît 163 fois dans l'intervalle [0 ; 0,5[.
   On note D_9 le neuvième décile de cette série statistique.
   Montrer que $D_9 \in [2,5\ ;\ 3[$.
4. En argumentant soigneusement la réponse, dire si pour la série observée au début, on peut affirmer avec un risque inférieur à 10 % que « la prairie est composée d'autant de fleurs de chaque variété ».

EXERCICE 10.6 (Adéquation à une loi binomiale)**.**
Une étude sur 1 000 familles de trois enfants a donné les résultats suivants :

Nombre de garçons i	0	1	2	3	Total
effectif e_i	110	384	382	124	1 000
fréquence f_i					1

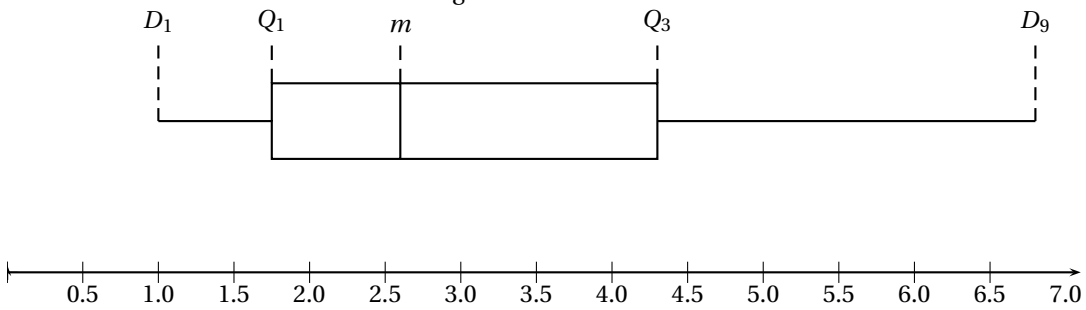
On se propose d'étudier la compatibilité de ses résultats avec l'hypothèse que la naissance d'une fille ou d'un garçon sont deux événements équiprobables.

1. Si cette hypothèse est correcte, la loi théorique du nombre de garçons dans une famille de trois enfants est une loi binomiale de paramètres 3 et 0,5.
   Compléter alors le tableau suivant :

Nombre de garçons i	0	1	2	3
probabilité p_i				

2. Pour une loi binomiale (et pour toute autre loi) on mesure la distance entre les fréquences observées et les probabilités théoriques par la formule :
$$\chi^2_{\text{obs}} = n\sum_{i=1}^{k}\frac{(f_i - p_i)^2}{p_i}$$
   Ici $n = 1\,000$; calculer la valeur de χ^2_{obs} (on pourra s'aider d'un tableur).
3. Une simulation sur 100 essais a donné les résultats résumés dans le diagramme de la figure 10.2 de la présente page.
   Indiquer une valeur approximative du neuvième décile de cette distribution.

FIGURE 10.2 – Diagramme en boite de l'exercice 10.6



4. La distribution observée est-elle en adéquation avec la loi $\mathscr{B}(3\,;0,5)$?
   L'hypothèse d'équiprobabilité pour la naissance d'une fille ou d'un garçon est-elle acceptable ?