ADÉQUATION À UNE LOI ÉQUIRÉPARTIE

1. Exemple introductif - Contexte

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 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, si on simule (par ordinateur par exemple) 10000 fois l'expérience consistant à lancer 50 fois une pièce de monnaie **bien équilibrée**, on réalisera qu'une proportion pas forcément négligeable de ces 10000 expériences a obtenu effectivement 30 fois PILE et 20 fois FACE. On peut donc considérer que l'événement "30 fois PILE et 20 fois FACE" otbenu lors de nos 50 lancers n'est pas si improbable que ça. On pourra ainsi décider si notre pièce est, elle aussi, bien équilibrée (au risque de commettre une erreur).
Nous sommes donc confrontés à plusieurs questions :

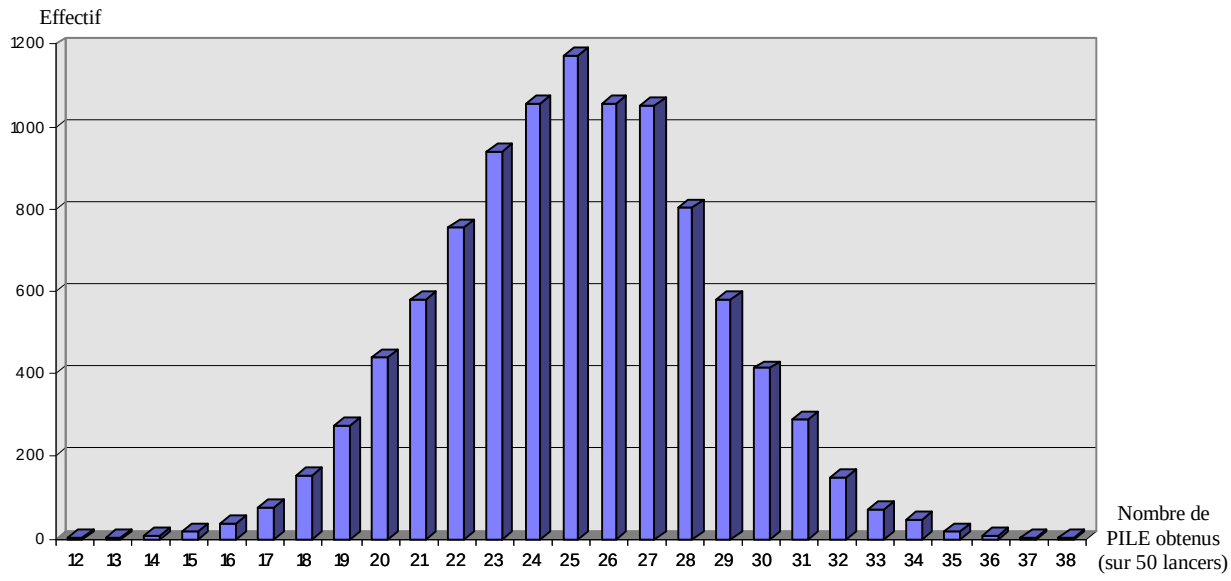
quels sont les critères de décisions ?

quelles sont les erreurs commises ?

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, 10000 fois l'expérience $\mathscr{E}$ pour une pièce bien équilibrée (hypothèse d'équirépartition) et on obtient les résultats suivants :

Nombre de "PILE" obtenus	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
effectifs	1	1	6	16	36	76	154	274	441	581	755	942	1057	1171	1055	1052	802	581	413	288	149	71	47	19	8	3	1

Par exemple, sur les 10000 expériences, 76 ont donné 17 PILE (et donc 33 FACE). On constate que 413 ont donné 30 PILE (et donc 20 FACE) ce qui représente tout de même 4,13% de l'effectif total.
Illustrons cette distribution avec le diagramme suivant :



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 de l'échantillonnage)

Nous allons calculer, par rapport à l'événement que nous avons observé ("30 fois PILE et 20 fois FACE") un premier indicateur D^2_{obs} :

$$D^2_{obs} = \left(\frac{3}{5}-\frac{1}{2}\right)^2 + \left(\frac{2}{5}-\frac{1}{2}\right)^2 = \frac{1}{50}$$

(D_{obs} représente la distance entre les fréquences observées et les fréquences théoriques)

Autrement dit, nous avons simplement : $5000\, D^2_{obs} = 100$

L'idée est de calculer la quantité $5000D^2$ pour chaque résultat de la simulation. Cela donne :

Nombre de "PILE" obtenus	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Valeurs de $5000D^2$	676	576	484	400	324	256	196	144	100	64	36	16	4	0	4	16	36	64	100	144	196	256	324	400	484	576	676
effectifs	1	1	6	16	36	76	154	274	441	581	755	942	1057	1171	1055	1052	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.

Réorganisons les données en faisant un tableau des valeurs de $5000D^2$ suivant 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	1171	3283	5277	6834	7996	8850	9412	9715	9862	9945	9980	9994	9998	10000

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 $5000\, D^2_{obs}$ 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 à rapport aux déciles :

> si $5000\, D^2_{obs} \leqslant D_9$ (neuvième décile), alors on considère que le modèle observé suit la loi d'équirépartition
>
> si $5000\, D^2_{obs} > D_9$ (9e décile), alors on considère que le modèle observé ne suit pas la loi d'équirépartition.

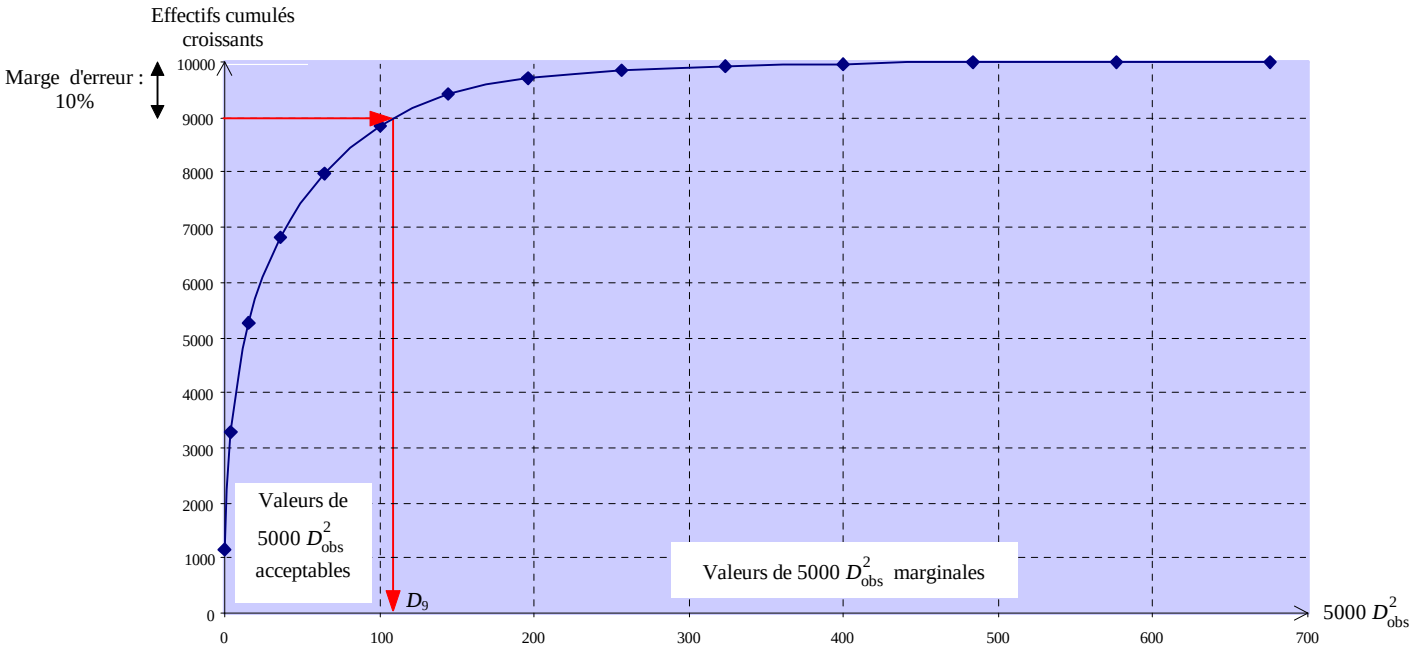
- si on se donne une autre marge d'erreur, par exemple 1%, on raisonne de même en comparant $5000\, D^2_{obs}$ au 99e centile.

Dans notre situation, calculons le 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. Parmi les valeurs de $5000D^2$, on prend celles qui sont inférieures ou égales à 9000. D'après le tableau, D_9 est compris entre 100 et 144. Si on veut le calculer plus finement, c'est possible en cherchant l'équation de la droite passant par les points de coordonnées (100 ; 8850) et B(144 ; 9412) :

On trouve : $$y = \frac{281}{22}x + \frac{83300}{11}$$

Lorsque $y = 9000$, on obtient : $D_9 = x = \dfrac{31400}{281} \simeq 112$ (à une unité près)

Si on ne souhaite pas une telle précision, il est bien plus commode de déterminer D_9 à l'aide du polygone des effectifs cumulés croissants :



$D_9 \simeq 110$

Quoiqu'il en soit, nous avons ici : $5000\, D_{obs}^2 \leqslant D_9$

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 :

$$5000\, D_{obs}^2 = 5000 \times \left[\left(\frac{32}{50} - \frac{1}{2} \right)^2 + \left(\frac{18}{50} - \frac{1}{2} \right)^2 \right] = 196$$

Cette valeur de $5000\, D_{obs}^2$ est trop marginale : $5000\, D_{obs}^2 > D_9$

Une telle pièce serait considérée comme non équilibrée avec une marge d'erreur de 10%. Par contre, si on se donne une marge d'erreur de 1%, on raisonne alors par rapport au $99^{ème}$ centile (qui vaut aux alentours de 300) alors cette pièce est considérée comme bien équilibrée. Cela peut paraître paradoxal mais s'explique ainsi : se donner 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 une tranche de 10%. Une marge d'erreur de 1% a donc tendance à valider notre observation comme étant presque toujours équirépartie ! Et si, cas extrême, on se donne une marge d'erreur de 0%, cela signifiera que toute observation tombant dans la plage de valeurs obtenues par simulation sera considérée comme suivant un modèle d'équirépartition (autrement dit, une marge d'erreur de 0% signifie qu'on ne prend pas le risque de dire que la pièce est déséquilibrée)

Réponses à quelques questions :

- les résultats peuvent-ils différer si on recommence une autre simulation ? Oui, car l'étendue peut être différente, les valeurs des déciles (ou centiles ou autres) également.

- Pourquoi avoir multiplié D^2 par 5000 ? Juste pour une meilleure lisibilité afin de travailler avec des entiers.
- Pourquoi ne rejette-t-on pas 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 considère 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 !

2. Généralisation

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

On suppose que l'on dispose de données simulées (un grand nombre de fois) sur un modèle théoriquement équiréparti et on étudie la série statistique des grandeurs D^2 obtenues.

Pour une marge d'erreur de 10%, on raisonne avec le 9^e décile D_9 de la série des D^2 :

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

Pour une marge d'erreur de 5%, on raisonne avec le 19^e vingtile V_{19} de la série des D^2 :

si $D^2_{\text{obs}} \leqslant V_{19}$, on considère que l'expérience observée est équirépartie avec une marge d'errreur de 5 %

Dans le cas contraire, on considère que l'expérience observée n'est pas équirépartie

Et ainsi de suite.

On n'en dira pas plus car aucune connaissance théorique n'est exigible sur ce sujet et l'étude d'exemples est, à ce niveau, plus formatrice.

3. Autre exemple

Une clinique fait des statistiques sur les naissances (naturelles et non provoquées) selon le jour de la semaine.

Sur 1000 naissances naturelles relevées, on obtient les résultats suivants :

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

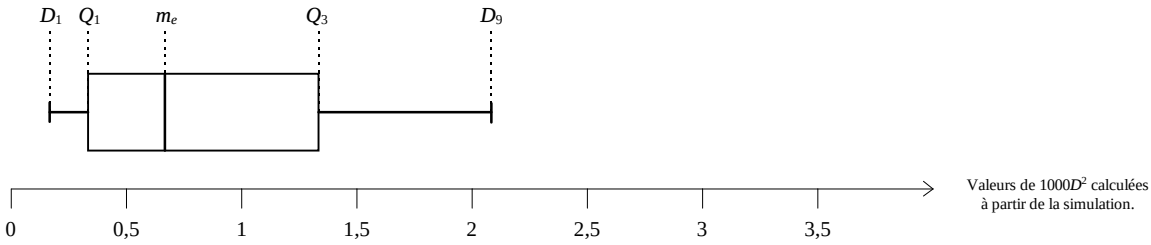
On s'intéresse à la validité de l'hypothèse "le nombre de naissance est indépendant du jour de la semaine".

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 : $$D^2_{\text{obs}} = \sum_{i=1}^{7}\left(f_i - \frac{1}{7}\right)^2$$

   Puis donner la valeur de $1000\,D^2_{\text{obs}}$ arrondie à 10^{-2} près. (On a multiplié par 1000 pour une meilleure lisibilité)

2. On simule sur un ordinateur 50000 séries de 1000 naissances équiréparties sur les sept jours de la semaine.
   Pour chacune de ces 5000 séries, l'ordinateur a calculé la valeur de $1000D^2$ (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 suivant :



   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 ?

Solution :

1. On calcule :

$$D_{\text{obs}}^2 = \left(\frac{146}{1000}-\frac{1}{7}\right)^2+\left(\frac{163}{1000}-\frac{1}{7}\right)^2+\left(\frac{158}{1000}-\frac{1}{7}\right)^2+2\times\left(\frac{156}{1000}-\frac{1}{7}\right)^2+\left(\frac{116}{1000}-\frac{1}{7}\right)^2+\left(\frac{105}{1000}-\frac{1}{7}\right)^2$$

$$D_{\text{obs}}^2 = 0{,}003145$$

$$1000\,D_{\text{obs}}^2 \simeq 3{,}14 \text{ à } 10^{-2} \text{ près}$$

2. On a largement : $1000\,D_{\text{obs}}^2 > D_9$

   On peut donc affirmer, avec un risque d'erreur de 10%, que les naissances dans cette clinique ne sont pas équiréparties sur la semaine (autrement dit, les variations des fréquences observées ne sont pas le seul résultat du phénomène de fluctuation de l'échantillonnage).

   Cela peut paraître surprenant sachant qu'il s'agit de naissances naturelles mais il y a certainement un phénomène psychosomatique qui est à prendre en considération chez les futures mamans : elles ne souhaiteraient pas accoucher le week-end (de peur d'être moins bien prises en charge), les contractions ne se déclencheraient qu'à partir du lundi ce qui expliquerait un pic de naissance le mardi... Autre explication : si les futures mamans restent plutôt au calme le week-end, elles ont aussi parfois beaucoup de tâches diverses à accomplir les jours de semaine ce qui peut favoriser l'apparition de contractions...