Exemple On lance un dé à six faces 1000 fois et on obtient les résultats suivants :

Chiffres obtenus	1	2	3	4	5	6
Effectifs	202	211	190	165	140	92
fréquences	0,202	0,211	0,19	0,165	0,140	0,092

Pour un dé équilibré, on s'attend à une loi équirépartie des fréquences :

Chiffres obtenus	1	2	3	4	5	6
fréquences	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$\frac{1}{6} \simeq 0,1666\ldots$

Le lancer d'un dé est une expérience aléatoire. On doit donc s'attendre à voir fluctuer la distribution des fréquences obtenues par rapport aux probabilités (distribution théorique).

Néanmoins, au vu des écarts observés ici, peut-on penser *raisonnablement* que le dé n'est pas correctement équilibré ?

Plus précisément, peut-on *quantifier* cet écart, et par la suite décider si celui-ci est *raisonnable* ou non ?

Position générale du problème : Dans une population on prélève un échantillon dont les effectifs des différentes modalités x_1, x_2, ...,x_k sont : n_1, n_2, ..., n_k.

Peut-on considérer que la distribution statistique observée dans cet échantillon est en adéquation avec une distribution théorique équirépartie ?

En d'autres termes, il s'agit de savoir si les écarts entre la distribution observée sur l'échantillon et une distribution théorique sont imputables aux fluctuations d'échantillonnage ou si ces écarts sont trop importants pour que l'on puisse accepter l'hypothèse :

> H : "L'échantillon est tiré d'une population caractérisée par une distribution équirépartie".

Soit n l'effectif total de l'échantillon, dans le cas d'une adéquation à une loi équiprobable la probabilité de chacune des k modalités est $p = \frac{1}{k}$, et les effectifs théoriques associés à chaque modalité sont $\frac{n}{k}$.

Si les valeurs n_i des effectifs observés lors de l'expérimentation sont "proches" des valeurs théoriques, il y a de "fortes chances" pour que la réponse au problème posé soit oui.

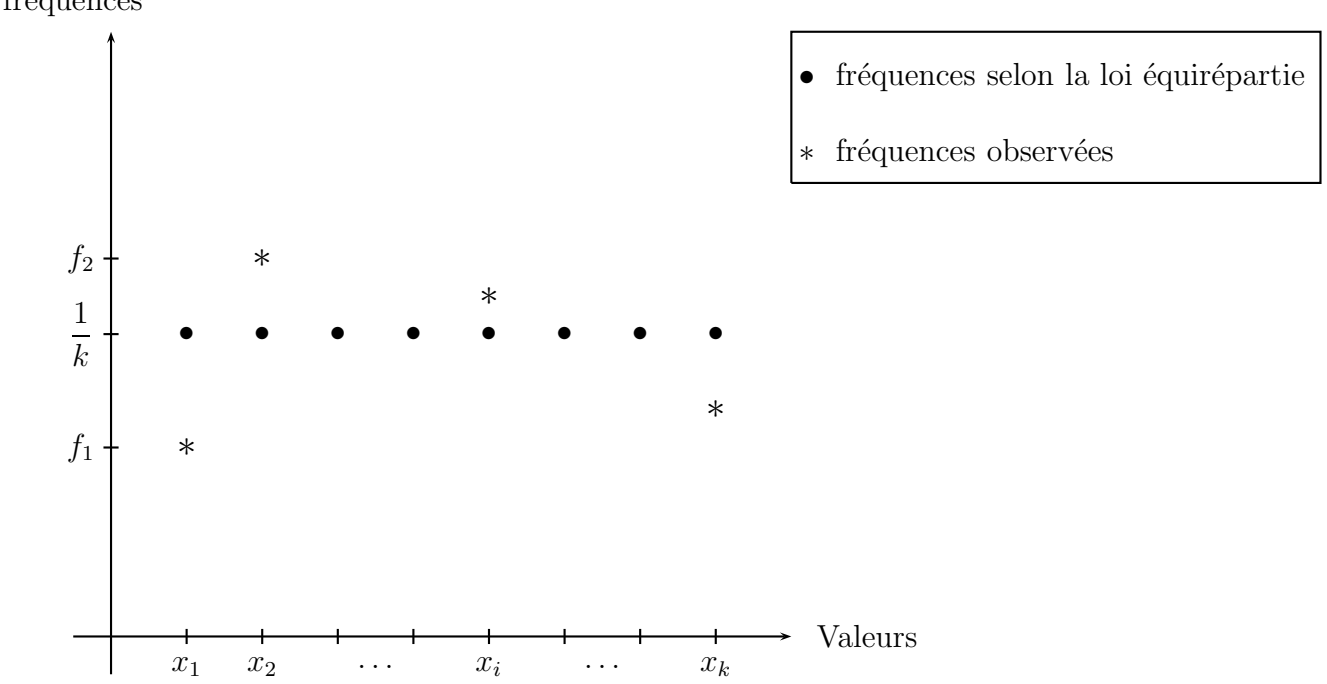
Comment quantifier cette proximité ?

Distribution observée

Valeur	x_1	x_2	...	x_i	...	x_k
Effectif	n_1	n_2	...	n_i	...	n_k
fréquence	f_1	f_2	...	$f_i = \frac{n_i}{n}$	...	f_k

Distribution suivant une loi équirépartie

Valeur	x_1	x_2	...	x_i	...	x_k
Effectif	$\frac{n}{k}$	$\frac{n}{k}$	...	$\frac{n}{k}$	...	$\frac{n}{k}$
fréquence	$\frac{1}{k}$	$\frac{1}{k}$	...	$\frac{1}{k}$	...	$\frac{1}{k}$



La notion de *proximité* entre les observations et la loi équirépartie peut se quantifier par le calcul de la distance :

$$d^2_{\text{obs}} = \left(f_1 - \frac{1}{k}\right)^2 + \left(f_2 - \frac{1}{k}\right)^2 + \cdots + \left(f_i - \frac{1}{k}\right)^2 + \cdots + \left(f_k - \frac{1}{k}\right)^2$$

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

Dans l'exemple du dé :

Chiffres obtenus	1	2	3	4	5	6
Effectifs	202	211	190	165	140	92
fréquences	0,202	0,211	0,19	0,165	0,140	0,092

on trouve

$$d^2_{\text{obs}} = \sum_{i=1}^{k} \left(f_i - \frac{1}{6}\right)^2 = \left(0,212 - \frac{1}{6}\right)^2 + \left(0,211 - \frac{1}{6}\right)^2 + \cdots + \left(0,092 - \frac{1}{6}\right)^2 \simeq 0,010$$

Interprétation de d^2_{obs} : La loi des grands nombres nous dit que plus n est grand (le nombre de répétitions de l'expérience, ici le nombre de lancers du dé), plus les fréquences observées se rapprochent de $\frac{1}{k}$, c'est-à-dire que, sous l'hypothèse H, d^2_{obs} tend vers 0 losrque n tend vers l'infini.

Si l'hypothèse H est fausse, alors les fréquences f_i vont tendre vers des valeurs dont au moins une est différente de $\frac{1}{k}$, et donc d^2_{obs} va tendre vers une valeur non nulle.

La question qui se pose alors est :

Est-ce que d^2_{obs} est assez proche de 0 pour considérer l'hypothèse H vraie ?

On prend cette décision en comparant la valeur de d^2_{obs} obtenue avec celles de N simulations aléatoires d'une loi équirépartie.

1. On calcule d^2_{obs}, qui mesure la distance entre la distribution des fréquences de notre échantillon avec la loi équirépartie.
2. On réalise N simulations d'échantillons de taille n de la loi équirépartie à k issues en calculant pour chacune la valeur de d^2 correspondante.
   On obtient ainsi une série de N valeurs d_1^2, d_2^2,...,d_n^2 dont on détermine le 9^e décile D_9 (valeur qui sépare les 90% de données les plus petites de la série des 10% les plus grandes).
3. En prenant le risque de rejeter à tort l'hypothèse d'équiprobabilité dans 10% des cas on convient alors, que :
   - Si $d^2_{\text{obs}} > D_9$, alors on peut rejeter, avec un risque d'erreur inférieur à 10%, l'hypothèse H, c'est-à-dire l'adéquation des données observées à une loi équirépartie.
   - Si $d^2_{\text{obs}} \leqslant D_9$, on ne peut pas, avec un risque d'erreur à 10%, rejeter l'adéquation des données à une série équirépartie.

A la suite d'un tel test, il y a donc quatre possibilités (avec un risque d'erreur de 10%) :
- L'hypothèse d'un modèle équiréparti est vraie et on opte pour l'adéquation à la fin du test.
- L'hypothèse d'un modèle équiréparti est fausse et on rejette le modèle équiréparti à la fin du test.

- L'hypothèse d'un modèle équiréparti est vraie et on la rejette à la fin du test.
- L'hypothèse d'un modèle équiréparti est fausse et on opte pour le modèle équiréparti à la fin du test.

Remarque : Si les simulations sont réalisés sur des échantillons de taille $n' \neq n$, on compare alors nd^2_{obs} au 9^e décile D_9 de la série des N valeurs $n'd^2$.

Exercice 1 *Les pics d'ozone*

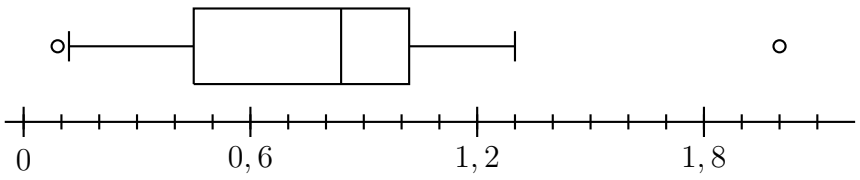
Dans le but de réduire les accidents mortels dus à l'alcool, la police d'une région a analysé les rapports de 175 accidents mortels dans lesquels le taux d'alcoolémie du conducteur était supérieur à la limite autorisée.

Le tableau suivant donne le répartition du nombre d'accidents en fonction du jour de la semaine :

Jour	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi	Dimanche
Nombre d'accidents	36	20	17	22	21	26	33

Il semble à première vue que la proportion d'accident mortels impliquant l'alcool soit plus importante le week-end.

On répète 1 000 simulation de 200 expériences modélisables par la loi équirépartie sur l'ensemble $\{1, 2, 3, 4?5, 6, 7\}$ et on obtient une série de 1 000 valeurs de $200d^2$ représentée par le diagramme en boîte suivant :



a) Peut-on, avec un risque d'erreur inférieur à 10%, rejeter l'adéquation à une loi équirépartie ?

b) Peut-onj conclure, au vu des résultats obtenus, qu'il y a plus d'accidents mortels dus à l'alcool le week-end ?

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

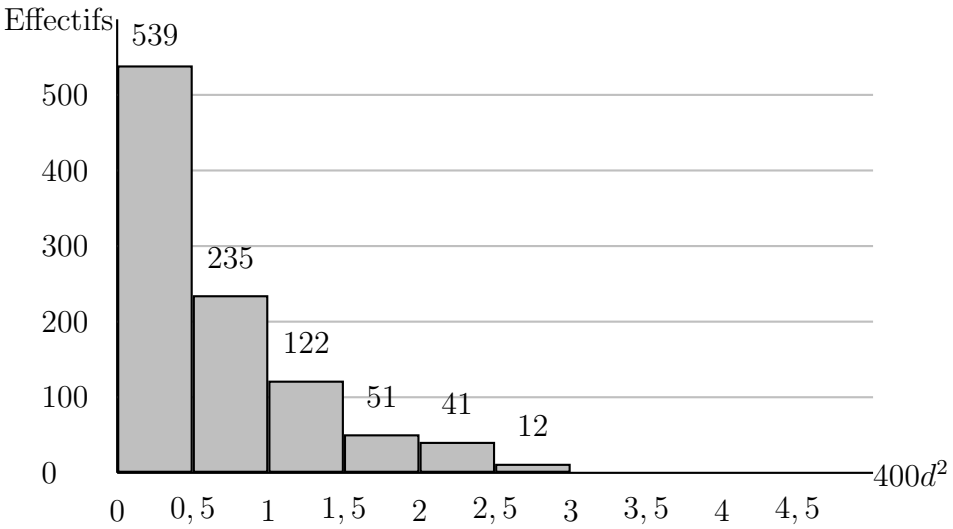
1. (a) Calculer les fréquences de prélèvement f_c d'une truite commune, f_s d'une truite saumonée et f_a d'une truite arc-en-ciel. On donnera les valeurs décimales exactes.

   (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 $400d^2_{\text{obs}}$ 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.

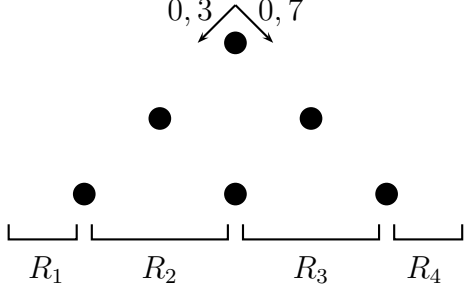


2. Déterminer une valeur approchée à 0,5 près par défaut, du neuvième décile D9 de cette série.
3. En argumentant soigneusement la réponse dire si on peut affirmer avec un risque d'erreur inférieur à 10 % que « le bassin contient autant de truites de chaque variété ».
4. On considère désormais que le 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.

1. On considère une planche à clous de ce type : On lance une boule B du haut de la planche, elle tombe alors dans l'un des quatre récipients notés R_1, R_2, R_3 ou R_4.
A chaque étape, la bille a une probabilité de 0,3 d'aller vers la gauche et 0,7 d'aller vers la droite (gauche et droite relatives à l'observateur).



On note p_1 la probabilité que la bille tombe dans le bac R_1 ou dans le bac R_3 et p_2 la probabilité qu'elle tombe dans le bac R_2 ou ou dans le bac R_4.
Que valent p_1 et p_2 ?

a) $p_1 = p_2 = 0,5$ b) $p_1 = 0,216$ et $p_2 = 0,784$ c) $p_1 = 0,468$ et $p_2 = 0,532$ d) $p_1 = 0,468$ et $p_2 = 0,432$

2. On a obtenu à l'aide d'un ordinateur les 1 000 premières décimales de π et on a compté le nombre d'occurences de chaque chiffre.

Chiffre	0	1	2	3	4	5	6	7	8	9
Nombre d'occurences	93	116	102	102	94	97	94	95	101	106

Avec un tableur, on a simulé 1 000 expériences de 1000 tirages d'un chiffre compris entre 0 et 9.
Pour chaque expérience, on a calculé $d^2 = \sum_{k=0}^{9}(f_k - 0,1)^2$, 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 le neuvième décile (d_1 et d_9), le premier et le troisième quartile (Q_1 et Q_3) et la médiane (M_e) :

$d_1 = 0,000\,422$ $Q_1 = 0,000\,582$ $M_2 = 0,000\,822$ $Q_3 = 0,001\,136$ $d_9 = 0,001\,45$

A. En effectuant le calcul de d^2 sur la série des 1 000 premières décimales de π, on obtient...
a) 0,000 456 b) 0,004 56 c) 0,000 314

B. 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.
Peut-il avec un risque d'erreur inférieur à 10% rejeter cette hypothèse ?
a) Oui a) Non c) Il ne peut rien dire