Chapitre 12 : Échantillonnage - simulation

Chapitre 12 : Échantillonnage - simulation

1 Échantillon pour une expérience à deux issues

1.1 Cas réel

Exemple

A partir de milliers de jeux de cartes mélangés, on a fabriqué une population de cartes noires (des piques et des trèfles) ou rouges (uniquement des cœurs) **d'effectif extrêmement grand**.

On a donc la proportion de cartes noires exactement égales à $p = \frac{2}{3}$ dans cette population, le reste des cartes étant rouges.

Population contenant une proportion p = 2/3 de cartes noires

On répète 30 fois le tirage d'une carte dans la population. Cela donne un échantillon de taille $n = 30$ cartes.

Chaque tirage d'une carte est une expérience aléatoire.

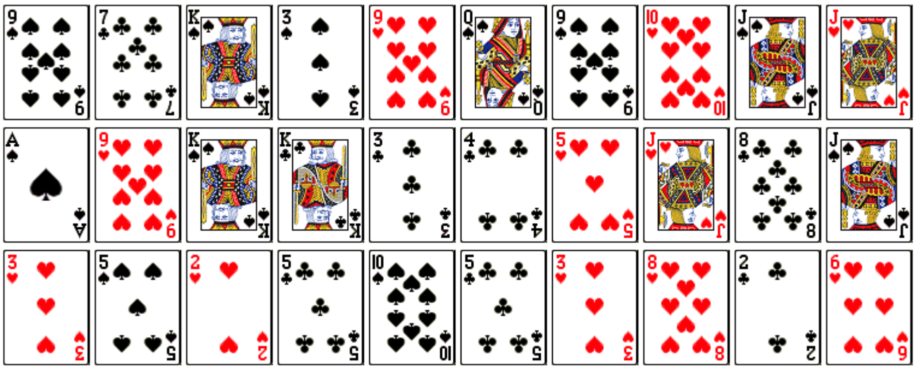
Comme la population est très grande, on peut considérer que les expériences sont indépendantes, c'est à dire que la couleur d'une carte obtenue à un tirage ne dépend pas des couleurs obtenues aux tirages précédents.

Définition

Lorsqu'on réalise plusieurs fois une même expérience de manière indépendante (c'est à dire que les différentes réalisations n'ont pas d'influence les unes sur les autres), l'ensemble des résultats obtenus est appelé échantillon. Le nombre de fois où l'expérience est réalisée est la **taille de l'échantillon**.

Exemples d'échantillons

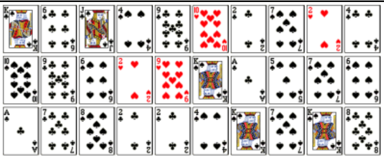
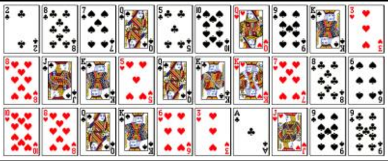
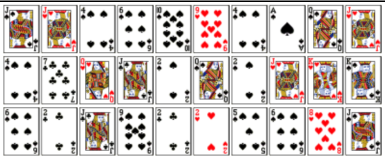
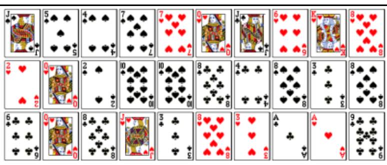
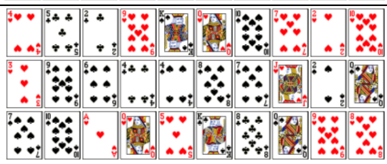
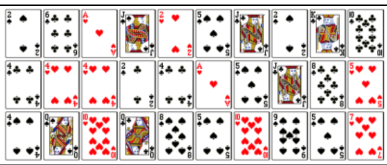
Voici un échantillon de taille $n = 30$. On remarque que la fréquence observée f_{obs} de cartes noires n'est pas obligatoirement égale à $\frac{2}{3}$. Dans cet échantillon on a **19 cartes noires** donc $f_{obs} = \frac{19}{30}$



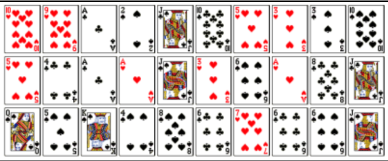
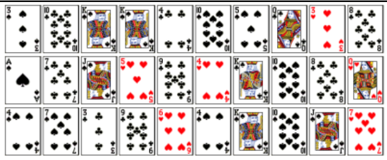
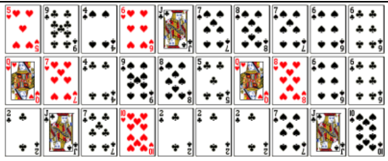
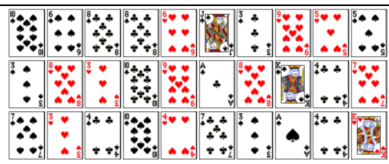
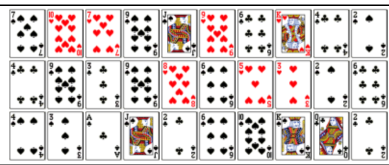
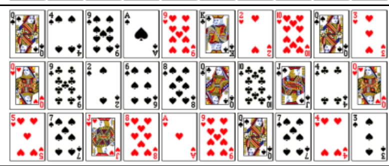
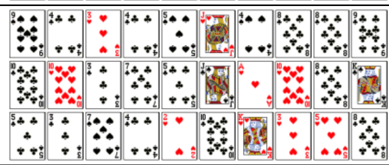
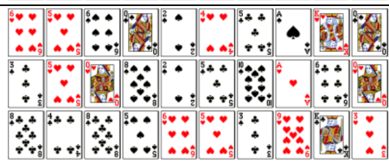
On peut obtenir plusieurs échantillons[1] de taille 30.

On note pour chacun la fréquence observée f_{obs} de cartes noires.

La population d'où proviennent les échantillons contient la proportion $p = \frac{2}{3} \approx 0{,}67$ de cartes noires.

Numéro de l'échantillon	Échantillon de taille $n = 30$	f_{obs}
1		$f_{obs} = \frac{17}{30} \approx 0{,}57$
2		$f_{obs} = \frac{16}{30} \approx 0{,}53$
3		$f_{obs} = \frac{26}{30} \approx 0{,}87$
4		$f_{obs} = \frac{19}{30} \approx 0{,}63$
5		$f_{obs} = \frac{18}{30} \approx 0{,}60$
6		$f_{obs} = \frac{22}{30} \approx 0{,}73$
7		$f_{obs} = \frac{22}{30} \approx 0{,}73$
8		$f_{obs} = \frac{17}{30} \approx 0{,}57$
9		$f_{obs} = \frac{21}{30} \approx 0{,}70$

[1] Site random.org dans la rubrique Games and Lotteries, Playing cards Schuffler

10		$f_{obs} = \frac{20}{30} \approx 0{,}67$
11		$f_{obs} = \frac{21}{30} \approx 0{,}70$
12		$f_{obs} = \frac{24}{30} \approx 0{,}80$
13		$f_{obs} = \frac{24}{30} \approx 0{,}80$
14		$f_{obs} = \frac{23}{30} \approx 0{,}77$
15		$f_{obs} = \frac{21}{30} \approx 0{,}70$
16		$f_{obs} = \frac{23}{30} \approx 0{,}77$
17		$f_{obs} = \frac{22}{30} \approx 0{,}73$
18		$f_{obs} = \frac{21}{30} \approx 0{,}70$
19		$f_{obs} = \frac{18}{30} \approx 0{,}60$
20		$f_{obs} = \frac{22}{30} \approx 0{,}73$

1.2 Simulation

Plutôt que de réaliser *en vrai* la population de cartes noires et rouges de très grand effectif et d'obtenir un ou plusieurs échantillons de 30 cartes réellement, on peut *simuler* de manière informatique ces expériences. Cela donne les mêmes résultats tout en étant beaucoup plus facile à réaliser.

On peut simuler informatiquement une expérience aléatoire à deux issues x_1 et x_2 de probabilités respectives p et $1-p$ en générant un réel aléatoire entre 0 et 1 et en considérant que :

- L'issue x_1 est réalisée si ce nombre aléatoire est inférieur ou égal à p.
- L'issue x_2 est réalisée si ce nombre aléatoire est strictement supérieur à p.

Reprise de l'exemple précédent : $p = \frac{2}{3}$

On pose x_1 l'issue "La carte est noire" et x_2 l'issue "La carte est rouge".

On réalise 20 tirages aléatoires d'un nombre réel sur l'intervalle $[0\ ;1]$

1.2.1 Sur la calculatrice

Sur la TI-83, touche $\boxed{\text{math}}$, menu PROB, on choisit NbrAléat.

On appuie 20 fois sur la touche $\boxed{\text{entrer}}$.

On obtient :

0,741 0,973 0,824, 0,980, 0,083, 0,411, 0,896, 0,509, 0,350, 0,286,
0,922, 0,519, 0,812, 0,091, 0,894, 0;020, 0,892, 0,962, 0,171, 0,036,
0,632 0,428, 0,252 0,233 0,675, 0,191, 0,428, 0,996, 0,167, 0,223.

- L'issue x_1 (carte noire) est réalisée si ce nombre aléatoire **est inférieur ou égal** à $p = \frac{2}{3} \approx 0{,}667$
- L'issue x_2 (carte rouge) est réalisée si ce nombre aléatoire est **strictement supérieur** à p.

Donc l'échantillon de taille $n = 30$ est :

0,741	0,973	0,824	0,980	0,083	0,411	0,896	0,509	0,350	0,286
0,922	0,519	0,812	0,091	0,894	0,020	0,892	0,962	0,171	0,036
0,632	0,428	0,252	0,233	0,675	0,191	0,428	0,996	0,167	0,223.

Sur cet échantillon on "observe" 22 "cartes noires".

On déduit que la fréquence observée est

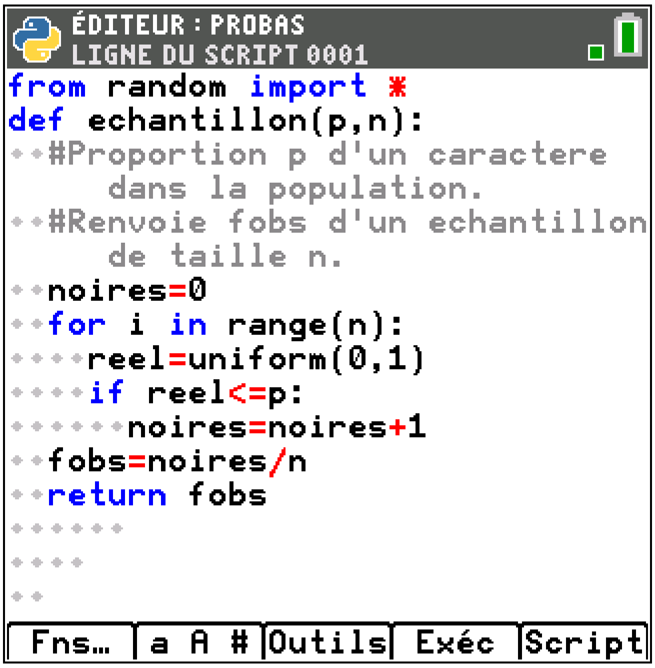
$$f_{obs} = \frac{22}{30} \approx 0{,}73$$

On peut recommencer plusieurs séries de 30 réels aléatoires sur $[0\,;1]$ pour simuler plusieurs échantillons. Mais cela peut devenir fastidieux si on veut par exemple simuler 20 échantillons pour obtenir 20 fréquences observées de cartes noires.

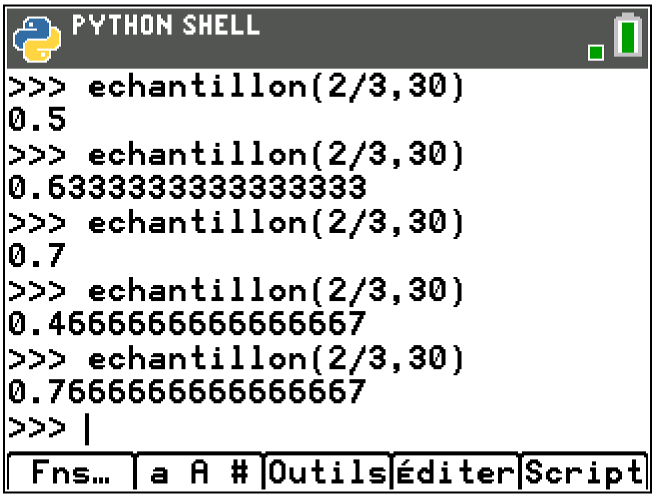
Dans ce cas, il est préférable de programmer une fonction en Python.

1.2.2 Avec un programme en Python

- Le programme a besoin de la fonction `uniform(0,1)` qui fait partie de la bibliothèque `random`. La fonction `uniform(0,1)` génère un réel entre 0 et 1 avec la même probabilité sur tout l'intervalle $[0\,;1]$. Si ce réel est inférieur à $\frac{2}{3}$ alors on considère que la carte est noire.



Ci-dessous, utilisation du programme pour simuler la production, dans une population où la proportion d'un caractère est $\frac{2}{3}$, de cinq échantillons de taille $n = 30$:



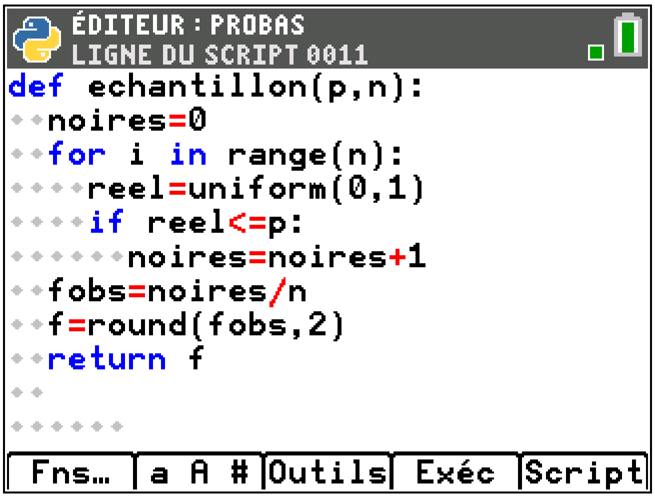
$$f_{obs} = 0{,}5 \ ; \quad f_{obs} \approx 0{,}63 \ ; \quad f_{obs} = 0{,}7 \ ; \quad f_{obs} = 0{,}47 \ ; \quad f_{obs} = 0{,}77 \ ; \quad \ldots$$

1.2.3 Amélioration en arrondissant

En Python, pour arrondir la variable `fobs` à 2 décimales et stocker le résultat dans la variable `f`, on utilise l'instruction :

```
f = round(fobs,2)
```

On obtient donc le programme :



1.2.4 Amélioration en stockant dans une liste

Une liste Python est un type de données qui se présente entre des crochets `[]`.

Les éléments contenus dans la liste sont séparés par une virgule. Par exemple :

```
ma_liste = [2, 35, 42, 39, 41]
```

- Pour créer une nouvelle liste, on écrit :

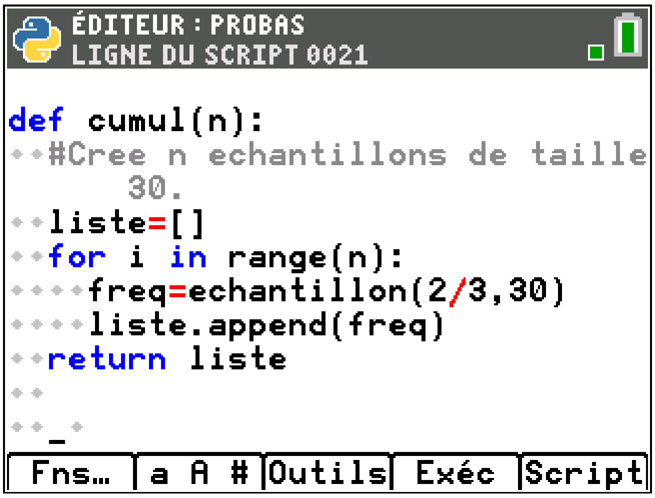
```
ma_liste = []
```

- Pour ajouter les éléments un par un, on utilise :

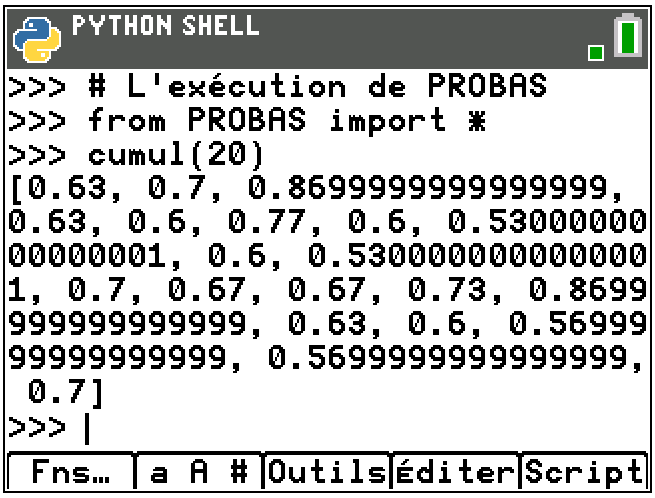
```
ma_liste.append(2)
ma_liste.append(35)
ma_liste.append(42)
```

- Pour appeler les éléments, un par un, on utilise leur rang qui commence à 0. Par exemple :

`ma_liste[0]` vaut 2 , `ma_liste[1]` vaut 35 , `ma_liste[2]` vaut 42 etc.



- En exécutant la fonction `cumul(20)` on obtient une liste de 20 échantillons :



On remarque que certains échantillons ont une fréquence mal arrondie à 2 décimales, mais cel n'est pas très gênant.

2 Estimation

2.1 Fluctuation d'échantillonnage

Les échantillons (obtenus par l'expérience ou simulés) de même taille ne sont pas identiques. Ce phénomène s'appelle **la fluctuation d'échantillonnage**.

Dans l'exemple précédent, c'est à cause de la fluctuation d'échantillonnage que la fréquence observée du caractère "carte noire" varie d'un échantillon à l'autre alors que la proportion $p = 2/3$ reste toujours la même dans la population.

2.2 Échantillon

Position du problème :

On connait la probabilité p d'apparition d'un caractère binaire[2] dans une population.

On prélève un échantillon de taille n.

Dans quel intervalle a-t-on une grande chance de trouver la fréquence observée ?

f_{obs} appartient, la plupart du temps à l'intervalle $\left[p-\frac{1}{\sqrt{n}}\ ;\ p+\frac{1}{\sqrt{n}}\right]$

Exemple

Dans le cas des échantillons de taille $n=30$ avec une proportion $p=\frac{2}{3}$ on trouve la plupart des fréquences observées dans l'intervalle $\left[\frac{2}{3}-\frac{1}{\sqrt{30}}\ ;\ \frac{2}{3}+\frac{1}{\sqrt{30}}\right]$ c'est à dire approximativement dans l'intervalle $[0{,}484\ ;\ 0{,}849]$

Exemple

Dans le cas des échantillons de taille $n=10000$ avec une proportion $p=\frac{2}{3}$ on trouve la plupart des fréquences observées dans l'intervalle $\left[\frac{2}{3}-\frac{1}{\sqrt{10000}}\ ;\ \frac{2}{3}+\frac{1}{\sqrt{10000}}\right]$ c'est à dire approximativement dans l'intervalle $[0{,}657\ ;\ 0{,}667]$.

2.3 Principe de l'estimation d'une probabilité

Position du problème :

<u>On ne connait pas</u> la probabilité p d'apparition d'un caractère binaire dans une population. On prélève un échantillon de taille n. On voudrait estimer p en procédant par échantillonnage. On observe dans l'échantillon de taille n la fréquence observée f_{obs}. Pour n grand (au moins 30), f_{obs} et p sont proches. Donc si on ne connait pas la valeur de p, alors f_{obs} en est une **estimation**. On utilise généralement plusieurs échantillons de même taille pour réaliser une bonne estimation.

[2] **Binaire :** soit l'individu possède ce caractère, soit il ne le possède pas. Par exemple soit la carte tirée est noire, soit elle ne l'est pas.

Exemple

Reprenons l'exemple de la population de cartes qui peuvent être noires ou non. Imaginons cette fois **qu'on ignore la proportion de cartes noires dans la population** très grande et qu'on n'a pas les moyens de vérifier la couleur de toutes les cartes de la population.

On va donc procéder par sondages.

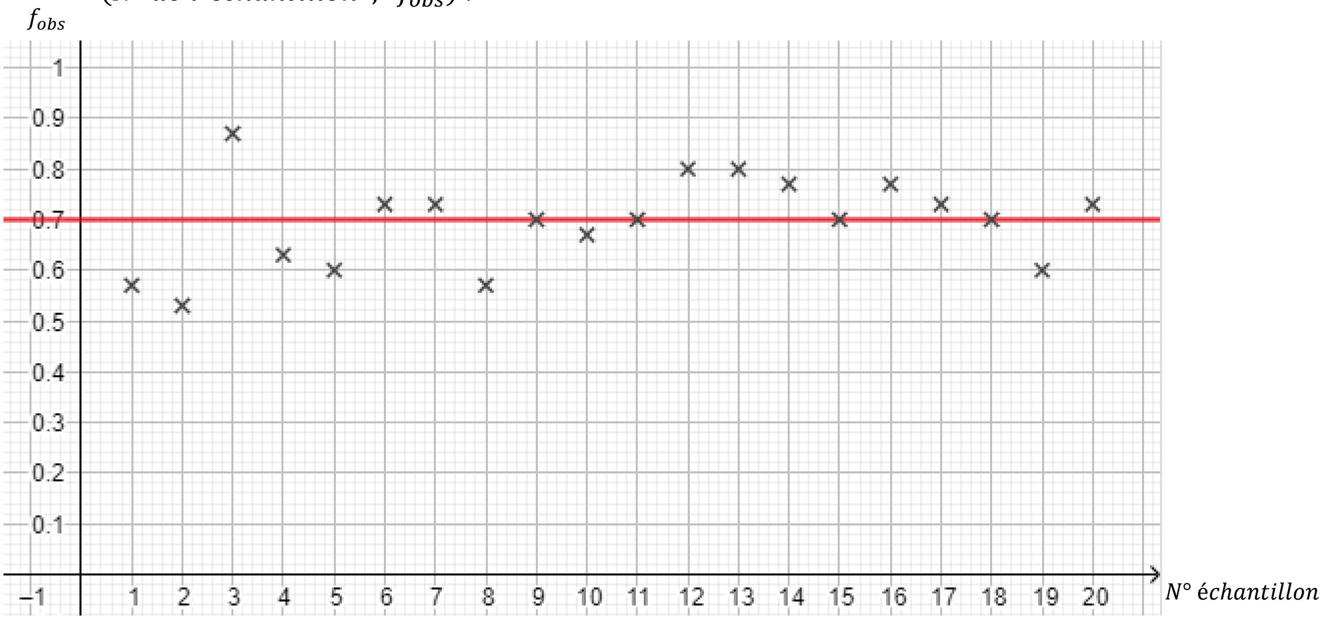
Les 20 échantillons de 30 cartes sont ceux du cas réel présenté au paragraphe 4.1.

- Rappelons les fréquences observées :

Numéro de l'échantillon	f_{obs}
1	$f_{obs} = \frac{17}{30} \approx 0{,}57$
2	$f_{obs} = \frac{16}{30} \approx 0{,}53$
3	$f_{obs} = \frac{26}{30} \approx 0{,}87$
4	$f_{obs} = \frac{19}{30} \approx 0{,}63$
5	$f_{obs} = \frac{18}{30} \approx 0{,}60$
6	$f_{obs} = \frac{22}{30} \approx 0{,}73$
7	$f_{obs} = \frac{22}{30} \approx 0{,}73$
8	$f_{obs} = \frac{17}{30} \approx 0{,}57$
9	$f_{obs} = \frac{21}{30} \approx 0{,}70$
10	$f_{obs} = \frac{20}{30} \approx 0{,}67$
11	$f_{obs} = \frac{21}{30} \approx 0{,}70$
12	$f_{obs} = \frac{24}{30} \approx 0{,}80$
13	$f_{obs} = \frac{24}{30} \approx 0{,}80$
14	$f_{obs} = \frac{23}{30} \approx 0{,}77$
15	$f_{obs} = \frac{21}{30} \approx 0{,}70$
16	$f_{obs} = \frac{23}{30} \approx 0{,}77$
17	$f_{obs} = \frac{22}{30} \approx 0{,}73$

18	$f_{obs} = \frac{21}{30} \approx 0{,}70$
19	$f_{obs} = \frac{18}{30} \approx 0{,}60$
20	$f_{obs} = \frac{22}{30} \approx 0{,}73$

- Pour mieux visualiser ces résultats, on place dans un repère les points de coordonnées $(N°\ de\ l'échantillon\ ;\ f_{obs})$:



- Enfin, on a tracé une droite rouge qui semble "au milieu" des points et qui donne donc une estimation de la proportion de cartes noires dans la population.
  On peut estimer qu'il y a $p = 0{,}7$ soit une proportion de 70 % de cartes noires dans la population. On remarque qu'à cause de la fluctuation d'échantillonnage certains échantillons (comme le n°2 et le n°3) ont une fréquence observée assez éloignée de la valeur "au milieu".