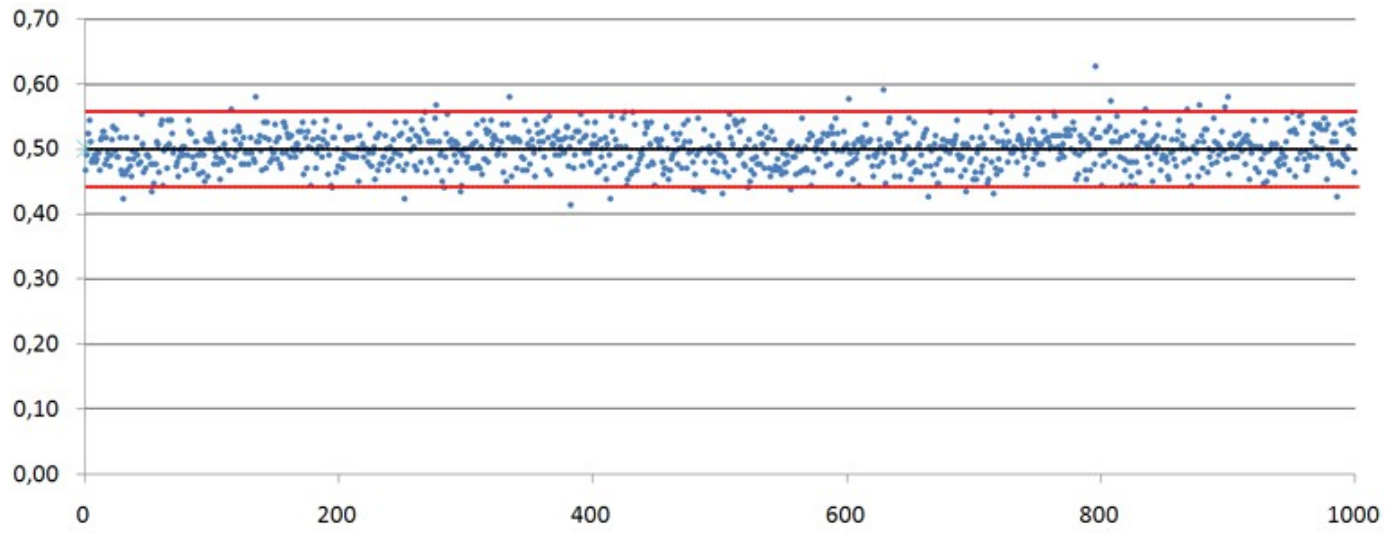
ÉCHANTILLONNAGE ET SIMULATION



I. Fluctuation d'échantillonnage

Tony « *l'embrouille* » vous propose de jouer à un jeu simple : il lance une pièce de monnaie, si le pile apparaît alors il vous donne 0,50 €, sinon c'est vous qui lui donnez 0,50 € ...

Vous décidez de jouer 50 parties.

Le malfrat lance donc sa pièce 50 fois : le pile apparaît avec une fréquence de 0,38 (soit 19 fois).

Est-il possible que la pièce soit truquée ?

Pour essayer de répondre à cette question, on se propose de faire l'expérience avec une pièce équilibrée.

On effectue 50 lancers d'une pièce.
La liste des 50 résultats « pile » ou « face » obtenus successivement en lançant 50 fois la pièce est un échantillon de taille $n=50$. On dit que cet échantillon relève du modèle de Bernoulli de probabilité 0,5.

Définitions :

1) En statistiques, un échantillon de taille n est la liste de n résultats obtenus par n répétitions indépendantes d'une même expérience aléatoire.
2) On dit qu'un échantillon relève du ***modèle de Bernoulli de probabilité p*** lorsqu'il n'y a que deux issues possibles : le « succès » avec la probabilité p et « l'échec » avec la probabilité $(1-p)$.

Si on recommence l'expérience plusieurs fois, les fréquences vont varier d'un échantillon à l'autre.
Ce phénomène est appelé ***fluctuation d'échantillonnage***.

Ainsi, la fréquence de la pièce de Tony « l'embrouille » ne semble pas étonnante : la face « pile » aurait pu apparaître avec une autre fréquence que 0,38.

Cependant, il faudrait faire plus d'expériences, afin d'observer comment varient les résultats.
Mais recommencer physiquement un grand nombre de fois une expérience est souvent très long et fastidieux. On peut donc utiliser un tableur, une calculatrice ou un algorithme pour simuler des expériences…

Simulation au tableur : 100 échantillons de taille 50

A l'aide d'un tableur (OpenOffice Calc, Excel) on peut simuler le lancer d'une pièce.
On associe au résultat « pile » la valeur 1, et au résultat « face » la valeur 0.

	A
1	Échantillon 1
2	0
3	1
4	0
5	0
6	0
7	1
8	1
9	0
10	1
11	0
12	0
13	1
48	0
49	0
50	1
51	0
52	Fréquence Pile
53	0,48

Exemple avec OpenOffice Calc :

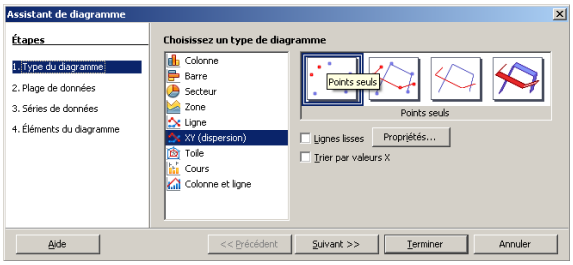
Dans la cellule A1, on écrit « échantillon 1 ».
Dans la cellule A2, on tape la formule : « =**ALEA.ENTRE.BORNES(**0 ;1**)** ».
Le tableur affiche alors un nombre aléatoire entier x tel que $0 \leqslant x \leqslant 1$,
c'est-à-dire 0 ou 1.
On étire cette formule jusqu'à la cellule A51.
Dans la cellule A53 on tape la formule : « =SOMME(A2:A51)/50 » pour avoir la fréquence d'apparition de « pile » dans ce premier échantillon de taille 50.

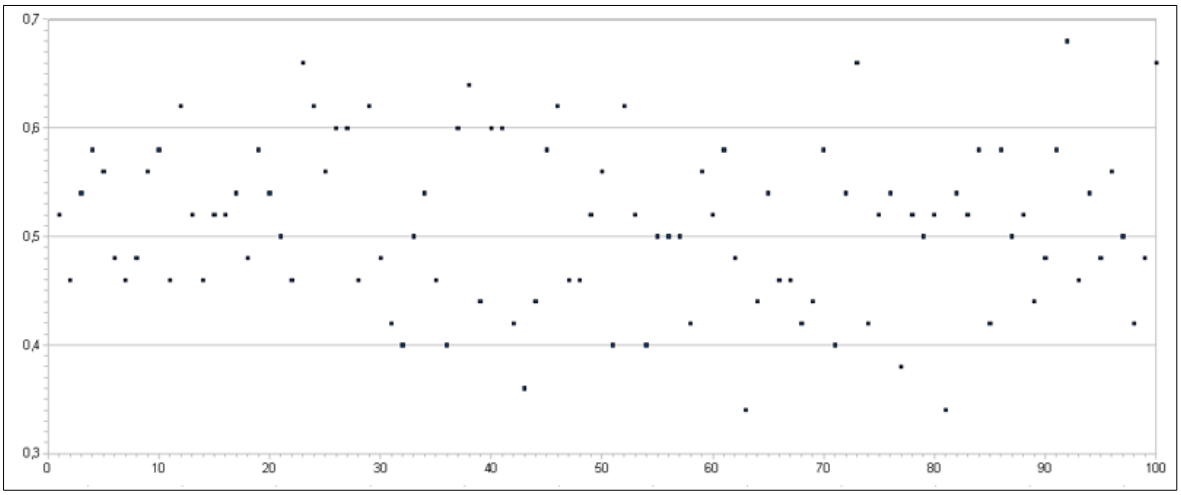
En appuyant sur *ctrl+shift+F9* (ou simplement *F9* pour Excel), le contenu de chaque cellule est mis à jour de façon aléatoire. On obtient ainsi un nouvel échantillon de taille 50.

Maintenant, pour obtenir 100 échantillons de taille 50, il suffit d'étirer la colonne A jusqu'à la colonne CV.

	A	B	C	D	CT	CU	CV
1	Echantillon 1	Echantillon 2	Echantillon 3	Echantillon 4	Echantillon 98	Echantillon 99	Echantillon 100
2	0	1	0	0	1	0	0
3	1	0	1	0	0	1	1
4	0	1	0	1	0	1	1
5	0	1	1	1	0	0	0
6	0	1	0	1	1	0	1
7	1	1	0	0	1	1	1
8	1	1	0	1	0	0	0
49	0	1	0	0	0	1	0
50	1	0	0	0	0	1	0
51	0	1	0	1	0	0	1
52	Fréquence Pile	Fréquence Pile	Fréquence Pile	Fréquence Pile	Fréquence Pile	Fréquence Pile	Fréquence Pile
53	0,48	0,54	0,6	0,54	0,4	0,44	0,38

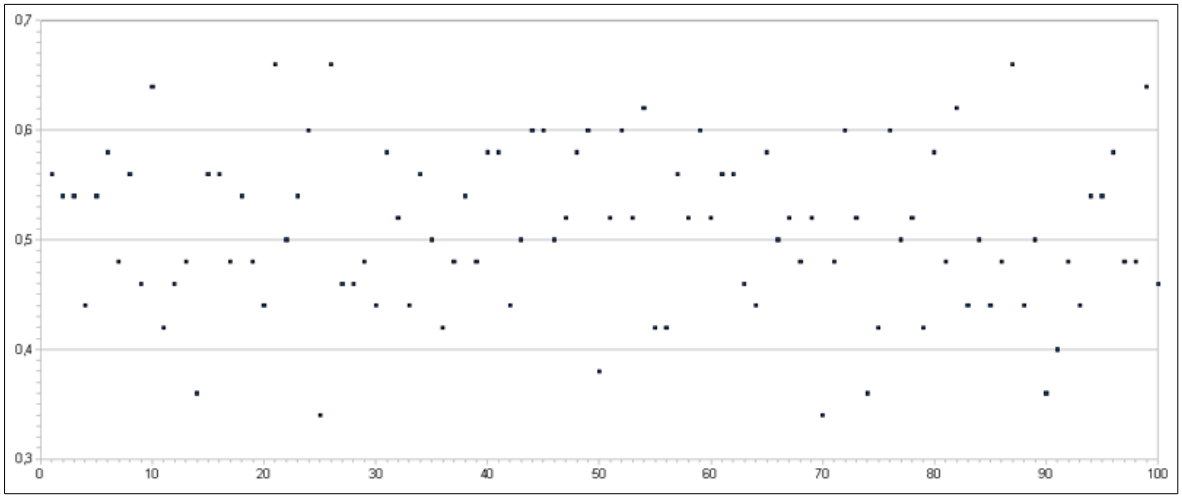
On peut alors insérer un nuage de points qui représente l'évolution des fréquences d'apparition du pile pour chaque échantillon. Pour cela, on sélectionne la plage A53:CV53, puis : « Insertion/Diagramme/XY (dispersion)/Points seuls » et « Terminer ».





Il est alors aisé de constater que sur ces 100 échantillons de taille 50, seuls 4 ont une fréquence d'apparition de pile strictement inférieure à 0,4.

En appuyant sur *ctrl+shift+F9*, on obtient un autre graphique :



Là aussi, sur ces 100 échantillons de taille 50, seuls 6 ont une fréquence d'apparition de pile strictement inférieure à 0,4.
On peut alors s'interroger sur la pièce de Tony… sans pour autant pouvoir affirmer qu'elle est truquée !
Il semble alors nécessaire d'avoir dans nos outils un résultat qui permette de répondre plus rigoureusement à la question posée.

PROPRIÉTÉ Si n est « assez grand » et p « ni trop petit ni trop grand » (on considère souvent que cela est vérifié lorsque $n \geqslant 30$, $n\,p \geqslant 5$ et $n(1-p) \geqslant 5$) :

pour environ 95 % des échantillons de taille n relevant du modèle de Bernoulli de probabilité p,

la fréquence d'apparition du « succès » appartient à l'intervalle :

Cet intervalle s'appelle ***intervalle de fluctuation au seuil de 95 %***.

Autrement dit :

Si, pour échantillon de taille n, on obtient une fréquence f de l'issue choisie, alors :

$$\mathrm{p}\left(f \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]\right) \approx 0{,}95 .$$

Ici, on a : $n =$ ..... ; $p =$ ..... ; $f =$ ..... .

L'intervalle de fluctuation au seuil de 95 % est :

Cela signifie que pour environ 95 % des échantillons de taille 50, la fréquence d'apparition de pile appartient à l'intervalle ..... .

Or, ...

On peut donc juger que la pièce de Tony .....

.....

II. Prise de décision à partir d'un échantillon

II.1 Tester la conformité d'un échantillon par rapport à la population

Deux entreprises A et B recrutent dans un bassin d'emploi où il y a autant de femmes que d'hommes, avec la contrainte du respect de la parité.
Dans l'entreprise A, il y a 100 employés dont 43 femmes.
Dans l'entreprise B, il y a 2500 employés dont 1145 femmes.
Quelle entreprise respecte le mieux la parité ?

1. a) Quel est le pourcentage de femmes dans l'entreprise A ?

b) Quel est le pourcentage de femmes dans l'entreprise B ?

c) En admettant que la parité c'est « 50 % de femmes, 50 % d'hommes », quelle entreprise respecte le mieux la parité ?

2. La parité, cela signifie que l'identité sexuelle n'intervient pas au niveau du recrutement, c'est-à-dire qu'au niveau du caractère homme ou femme, les résultats observés pourraient être obtenus par choix au hasard des individus dans la population. Dans ce cadre, l'entreprise A est assimilable à un échantillon de taille 100 du modèle de Bernoulli (avec $p=0{,}5$), et l'entreprise B à un échantillon de taille 2500 du même modèle.

a) Donner les intervalles de fluctuation au seuil de 95 % pour chaque échantillon.

b) Conclure : quelle entreprise respecte le mieux la parité ?

II.2 Conjecturer une proportion et valider/invalider cette hypothèse

Sur le site internet du Louvre[1], on peut lire le nombre de visiteurs en 2012 : 9,7 millions, dont 40 % de jeunes de moins de 26 ans et 52 % de moins de 30 ans. Il est également indiqué que « globalement, la fréquentation des visiteurs étrangers a progressé en 2012 (+11 %, soit, en volume, un surcroît d'environ 660 000 visites) avec 69 % de visiteurs étrangers ».
Une nouvelle exposition, plutôt destinée aux jeunes, est mise en place : sur trois jours, un sondage est effectué auprès de 387 visiteurs. Ont été comptabilisés 173 jeunes de moins de 26 ans, soit 44,7 % environ.

Cette exposition a-t-elle eu un impact sur la fréquentation du Louvre par les jeunes de moins de 26 ans ?

1 http://www.louvre.fr/sites/default/files/medias/medias_fichiers/fichiers/pdf/louvre-rapport-activites-2012.pdf