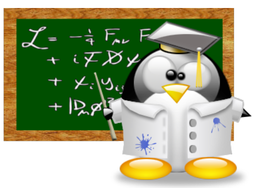
Stéphane PASQUET, 12 mai 2017

Sommaire



Prérequis

- Notions d'intervalles
- Statistiques (notion de fréquence)
- Probabilités de 3e

Définitions *Expérience aléatoire et échantillon*

Une `expérience aléatoire` est un acte comportant plusieurs issues possibles (connues) mais imprévisibles à l'avance.

Un `échantillon de taille n` est constitué des résultats de n répétitions indépendantes de la même expérience aléatoire.

Exemples

- Une expérience aléatoire peut être :
  - $\rightarrow$ le lancer d'un dé à 6 faces ;
  - $\rightarrow$ le lancer d'une pièce de monnaie ;
  - $\rightarrow$ un sondage d'opinion.
- Un échantillon peut être :
  - $\rightarrow$ les résultats de 100 lancers de dé lorsque l'on regarde si l'on obtient un multiple de 3 ; dans ce cas, la taille de l'échantillon est $n = 100$.
  - $\rightarrow$ les résultats d'un sondage sur 100 000 personnes pour savoir s'ils aiment ou s'ils n'aiment pas un certain parfum ; dans ce cas, la taille de l'échantillon est $n = 100\,000$.

Définition *Fluctuation d'échantillonnage*

Imaginons maintenant que l'on réalise un sondage dans les villes A et B. Prenons deux échantillons dans chacune des deux villes, de tailles différentes.
Les résultats seront très certainement différents d'une ville à l'autre et les fréquences d'apparition d'un caractère donné (par exemple, le pourcentage de personnes ayant répondu « oui » à une question) seront différentes.

On appelle `fluctuation d'échantillonnage` les variations de ces fréquences.

Pour comprendre ceci, nous allons réaliser une expérience à l'aide d'un tableur.

Simulation d'un lancer de dé cubique

Ouvrons un tableur et entrons dans la cellule `A1` la formule suivante :

	A	B	C
1	=ENT(1+6*ALEA())		
2			

- La fonction `ALEA()` permet de calculer un nombre aléatoire compris entre 0 (compris) et 1 (non compris) ;
- `6*ALEA()` affichera donc un nombre dans l'intervalle $[0\,;6[$;
- `1+6*ALEA()` affichera donc un nombre dans l'intervalle $[1\,;7[$;

- La fonction `ENT` calcule la partie entière d'un nombre ; donc `ENT(1+6*ALEA())` affichera un nombre entier parmi : 1, 2, 3, 4, 5 et 6.

Maintenant, copions cette formule de `A2` à `A100` afin d'obtenir, au final, 100 nombres entiers compris entre 1 et 6.
Dans la cellule `A101`, entrons la formule :

```
=NB.SI(A1:A100;1)
```

Elle affichera le nombre de « 1 » qui apparaissent dans la colonne `A`.
Ensuite, entrons dans la cellule `A102` la formule :

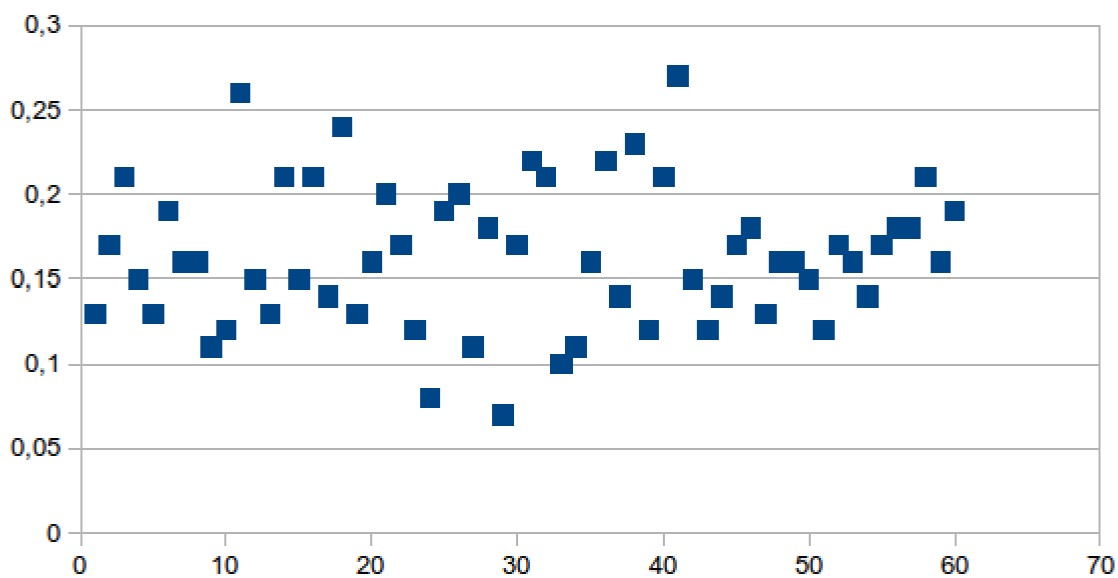
```
=A101/100
```

pour afficher la fréquence d'apparition du nombre « 1 » dans la colonne `A`.

Copions alors la colonne `A` dans les colonnes de `B` à `BG` afin d'obtenir au final 60 colonnes de nombres.
Ceci simule 60 échantillons de 100 lancers d'un dé cubique (60 échantillons de taille $n = 100$).

On construit alors un nuage de points représentant les différentes fréquences obtenues.

En appuyant sur la touche `[F9]`, on rafraîchit les calculs et on obtient un autre graphique. Voici un des graphiques obtenus :



La fréquence d'apparition théorique du nombre « 1 » est la probabilité d'obtenir ce nombre : $p = \dfrac{1}{6} \approx 0,17$.

On s'aperçoit alors sur le graphique que les fréquences semblent se regrouper autour de cette valeur.
On peut donc imaginer que les fréquences d'apparition du nombre « 1 » vont se trouver dans un intervalle centré en p : un intervalle de la forme $I = [p - \delta\,; p + \delta]$.

Nous allons prendre $\delta = \dfrac{1}{\sqrt{n}} = \dfrac{1}{\sqrt{100}} = 0,1$. Alors, $I \approx [0,066\,; 0,267]$.

Pour connaître le pourcentage de fréquences comprises dans cet intervalle, insérons dans la cellule `A103` la formule :

```
=SI(A102<=0,266;SI(A102>0,067;1;0);0)
```

qui affiche « 1 » si tel est le cas, et « 0 » sinon. Ensuite, copions-la sur toute la ligne 103 jusqu'à `BH103`.
Dans la cellule `A104`, insérons la formule :

```
=SOMME(A103:BH103)/60
```

puis formatons cette cellule de sorte à ce qu'elle affiche un pourcentage (en cliquant sur l'icône « % ») : ceci affichera le pourcentage de fréquences d'apparition du « 1 » qui sont dans I.

En appuyant sur la touche `[F9]` plusieurs fois, je vois les pourcentages suivants :

100% ; 96,67% ; 98,33% ; 100% ; 100% ; 98,33%.

En utilisant donc cette valeur de δ, nous sommes *a priori* assurés qu'au moins 95% des fréquences se trouvent dans I.

Définition — *Intervalle de fluctuation au seuil de 95%*

Un intervalle de fluctuation au seuil de 95%, relatif aux échantillons de taille n, est un intervalle centré en p qui contient la fréquence observée f dans un échantillon de taille n avec une probabilité égale à 0,95.

Propriété

Considérons un échantillon de taille $n \geqslant 25$ et posons p la probabilité d'obtenir un résultat précis. Notons f la fréquence d'apparition de ce résultat dans l'échantillon.
Si $0,2 \leqslant p \leqslant 0,8$, alors $f \in \left[p - \frac{1}{\sqrt{n}}\,; p + \frac{1}{\sqrt{n}}\right]$ avec une probabilité d'environ 0,95.

Méthode — *Prise de décision*

Cette dernière propriété peut servir à contrôler, par exemple, si un dé est truqué.

Pour tester un dé, lançons-le n fois, $n \geqslant 25$, et observons la fréquence d'apparition f d'une face (n'importe laquelle) sur ces n lancers.
Ensuite, calculons l'intervalle $I = \left[p - \frac{1}{\sqrt{n}}\,; p + \frac{1}{\sqrt{n}}\right]$, où p est la probabilité attendue.

- Si $f \notin I$, alors on peut rejeter l'hypothèse selon laquelle le dé est équilibré avec une probabilité de se tromper de 5% ;
- sinon, on accepte l'hypothèse selon laquelle le dé est équilibré.

Définition — *Intervalle de confiance au seuil de 95%*

Un `intervalle de confiance` permet d'avoir un intervalle où se situe la probabilité p avec une probabilité de 0,95%.

Propriété

Considérons un échantillon de taille $n \geqslant 25$ tel que $0,2 \leqslant f \leqslant 0,8$, où f est la fréquence d'apparition d'un caractère donné dans l'échantillon considéré.
Alors, la probabilité p du caractère est telle que $p \in \left[f - \dfrac{1}{\sqrt{n}}\,;f + \dfrac{1}{\sqrt{n}}\right]$ avec une probabilité égale à 0,95.

Remarque

Les intervalles de confiance sont très souvent utilisés dans les sondages (on parle de « fourchette de sondages »).

Exemple

Une semaine avant des élections confrontant les candidats Gérard Padbol et Lydie Adupot, un sondage a été réalisé auprès de 873 personnes. Les intentions de vôtes ont été synthétisés dans le tableau suivant :

G. Padbol	46%
L. Adupot	54%

Un intervalle de confiance au seuil de 95% du candidat G. Padbol est :

$$I_1 = \left[0,46 - \frac{1}{\sqrt{873}}\,;0,46 + \frac{1}{\sqrt{873}}\right] \approx [0,426\,;0,494].$$

Il faut qu'au moins 50% des vôtes soient pour un candidat pour qu'il gagne les élections.
On peut donc dire que G. Padbol ne sera pas élu (au risque de se tromper de 5%) car son intervalle de confiance ne dépasse pas 0,5.

Dans le calcul des intervalles de fluctuation et de confiance,

- on arrondira toujours la borne inférieure **par défaut** ;
- on arrondira toujours la borne supérieure **par excès**.