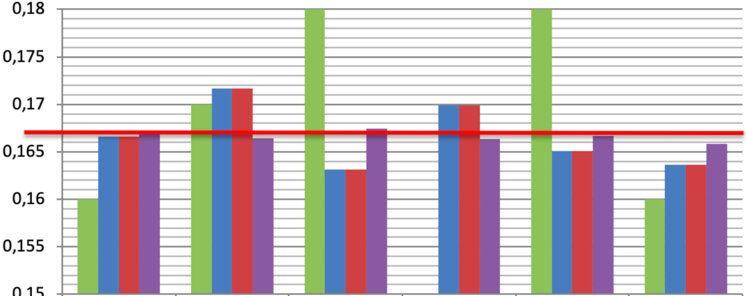
Échantillonnage

MatheX

19 mai 2020

Fréquence des issues en fonction du nombre d'essais

p=1/6



	1	2	3	4	5	6
n=100	0,16	0,17	0,23	0,1	0,18	0,16
n=1 000	0,1666	0,1717	0,1631	0,1699	0,1651	0,1636
n=10 000	0,1666	0,1717	0,1631	0,1699	0,1651	0,1636
n=100 000	0,16723	0,16645	0,16748	0,16634	0,16665	0,16586

Définition 1 : (échantillon, fréquence)

Soit E une expérience aléatoire et A un évènement de E

Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de E.

La fréquence observée de A est la proportion de réalisation de l'évènement A dans l'échantillon :

$$\boxed{f = \frac{k}{n}}$$

avec k le nombre de réalisation de A parmi les n répétitions de E

NB : en statistique, on parle de la fréquence obsservée d'un caractère dans une population

Exemple :

E : "lancer d'un dé" $\Omega = \{1, 2, 3, 4, 5, 6\}$
A :"je tombe sur 6"

Échantillon de taille $n = 10$:
échantillon 1 : "1 4 1 2 6 3 5 2 4 6"
échantillon 2 : "2 6 1 6 3 5 2 4 6 1"

Fréquence observée de A :

échantillon 1 : $k_1 = 2 \Rightarrow f_1 = \dfrac{k_1}{n} = \dfrac{2}{10} = 0,2$

échantillon 1 : $k_2 = 3 \Rightarrow f_2 = \dfrac{k_2}{n} = \dfrac{3}{10} = 0,3$

Exemple 2 :

Je fais un sondage pour connaître les intentions de vote pour le représentant des lycéens. Comme je ne peux pas interroger tout le monde, je prends un échantillon représentatif.

E : "choisir un élève" Ω ={élèves du lycée}
A :"candidat A" B :"candidat B" C :"candidat C"

Échantillon de taille $n = 10$:
échantillon 1 : "A A B C A A C B A C"
échantillon 2 : "C C A B A C B A C C"

Fréquence observée de A, B et C :
échantillon 1 : $f_A = 0,5$ $f_B = 0,2$ $f_C = 0,3$
échantillon 2 : $f_A = 0,3$ $f_B = 0,2$ $f_C = 0,5$

Définition 2 : (fluctuation d'échantillonnage)

Les fréquences observées varient selon les échantillons, on appelle ce phénomène la fluctuation d'échantillonnage .

Exemple :

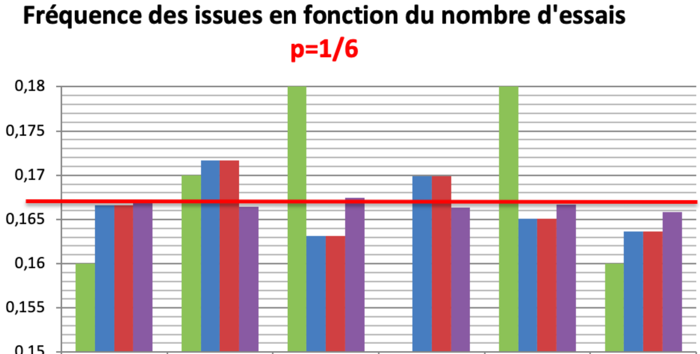
Variation des fréquences des échantillons 1 et 2 dans les deux exemples précédent.

Propriété 1 : (loi des grands nombre)

Dans la grande majorité des cas, plus la taille de l'échantillon augmente plus les fréquences observées des échantillon seront proches de la probabilité et moins elles seront dispersées :

$$\boxed{\text{pour n grand : } f \simeq P}$$

Exemple :



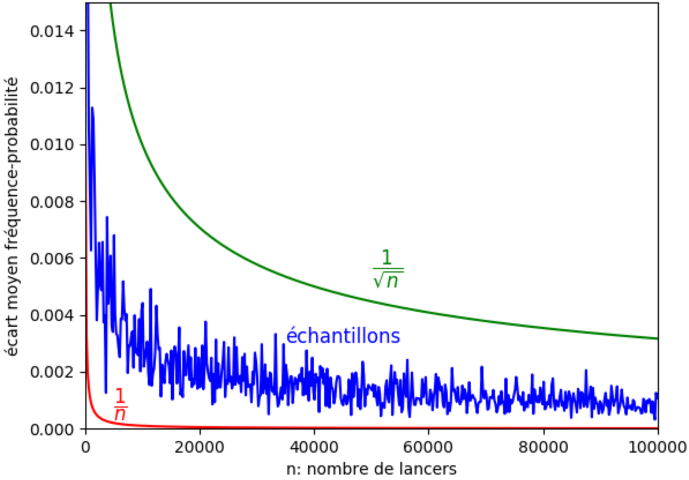
	1	2	3	4	5	6
n=100	0,16	0,17	0,23	0,1	0,18	0,16
n=1 000	0,1666	0,1717	0,1631	0,1699	0,1651	0,1636
n=10 000	0,1666	0,1717	0,1631	0,1699	0,1651	0,1636
n=100 000	0,16723	0,16645	0,16748	0,16634	0,16665	0,16586

Propriété 2 : (écart entre f et P)

Pour une taille d'échantillon n assez grande, dans la grande majorité des cas, l'écart entre f et P est majoré par $\dfrac{1}{\sqrt{n}}$:

$$\boxed{\left| f - P \right| < \frac{1}{\sqrt{n}}}$$ pour n grand

Exemple :


0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
écart moyen fréquence-probabilité
0
20000
40000
60000
80000
100000
n: nombre de lancers
$\frac{1}{\sqrt{n}}$
échantillons
$\frac{1}{n}$

Méthode 1 : (estimation)

Pour estimer une probabilité (ou la proportion d'un caractère dans une population), ont peut :

∘ prendre un échantillon de taille n
∘ calculer sa fréquence
∘ estimer la probabilité à la valeur de la fréquence observée

NB :
On n'est pas sûr mais dans la grande majorité des cas, l'erreur est majorée par $\dfrac{1}{\sqrt{n}}$

Exemple :

Je lance une pièce 100 fois et je tombe sur face 30 fois.

Je peux dire que la probabilité de face est 0,3

Je peux dire avec assez de certitude que cette pièce est truquée :

$$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0,1 \quad \text{et} \quad \left| f - P \right| = \left| 0,3 - 0,5 \right| = 0,2$$

Si la probabilité de face était 0,45 on pourrait dire avec assez de certitude que la pièce n'est pas truquée.

Dans tous les cas, on peut reprendre d'autres échantillons ou augmenter la taille de l'échantillon pour réduire le risque d'erreur de notre estimation.

Exemple 2 :

Je fais un sondage pour connaître les intentions de vote pour le représentant des lycéens.
Je prends un échantillon représentatif de la population et je sonde ces élèves sur le caractère "intention de vote" :
A :"candidat A" B :"candidat B" C :"candidat C"

J'interroge 100 élèves et je trouve :
$f_A = 0,29 \quad f_B = 0,2 \quad f_C = 0,51$

Je peux estimer les intentions de vote de tout le lycée :
$P_A = 0,29 \quad P_B = 0,2 \quad P_C = 0,51$

Plus précisément, je peux dire avec assez de certitude que :
$0,19 < P_A < 0,39 \quad 0,1 < P_B < 0,3 \quad 0,41 < P_C < 0,61$