ECHANTILLONNAGE

Objectifs *: Utilisation de la loi binomiale pour une prise de décision à partir d'une fréquence. Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion.*

Rappels de 2nd :

Propriété admise : Soit un caractère dont la proportion dans la population donnée est *p*.
Lorsque $n \geq 25$ et $0,2 \leq p \leq 0,8$, il y a 95% des échantillons de taille *n* issus de cette population qui sont tels que la fréquence *f* du caractère dans l'échantillon appartient à un intervalle centré en *p* de la forme $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}}\right]$.

Remarque : Cet intervalle $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}}\right]$ s'appelle intervalle de fluctuation au seuil de 95 %, ou au risque de 5%.

Application : Prise de décision à partir d'un échantillon.

Pour apprécier si une fréquence observée *f* sur un échantillon de taille *n* est compatible ou non avec un modèle de Bernoulli de probabilité *p*, on teste l'appartenance de *f* à l'intervalle de fluctuation $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}}\right]$ au seuil de 95%.

- Si *f* n'est pas dans l'intervalle de fluctuation, alors on peut rejeter l'hypothèse que l'échantillon soit compatible avec le modèle.
- Si *f* est dans l'intervalle de fluctuation, alors on ne peut pas rejeter l'hypothèse que l'échantillon soit compatible avec le modèle.

Quelle que soit la décision prise, il y a toujours le risque que ce ne soit pas la bonne décision dans 5% des cas.

I- Intervalle de fluctuation avec la loi binomiale

La loi binomiale permet de calculer très exactement les probabilités des différentes fréquences observables dans un échantillon de taille *n*, à savoir les valeurs $\frac{k}{n}$, avec $0 \leq k \leq n$, probabilités qui peuvent être représentées à l'aide d'un diagramme en bâtons. On peut également calculer, à l'aide d'un tableur, les probabilités (cumulées) des événements suivants : « *La fréquence observée dans l'échantillon prélevé de taille n est comprise entre 0 et* $\frac{k}{n}$ », événement qu'on peut aussi écrire $\left\{0 \leq F \leq \frac{k}{n}\right\}$, si *F* désigne la variable aléatoire qui à tout échantillon de taille *n* associe la fréquence observée dans l'échantillon prélevé.

En faisant varier les paramètres *n* et *p*, on observe que le diagramme n'est pas toujours symétrique, et non exactement centré sur *p*. Par ailleurs, le caractère discret de la loi binomiale fait qu'il n'est pas possible de déterminer précisément un intervalle où la fréquence observée se situe avec une probabilité égale à 0,95. La définition donnée en seconde pour l'intervalle de fluctuation suppose en effet, de manière implicite, que la fréquence suit une loi continue.
Pour ces différentes raisons, on est amené à construire un intervalle qui approxime l'intervalle de fluctuation défini plus haut en adoptant la définition suivante :

L'intervalle de fluctuation au seuil de 95 % d'une fréquence *F*, correspondant à la réalisation, sur un échantillon aléatoire de taille *n*, de la variable aléatoire *X* égale à *nF* et de loi binomiale de paramètres *n* et *p*, est l'intervalle $\left[\frac{a}{n}, \frac{b}{n}\right]$ défini par le système de conditions suivant :

***a* est le plus grand entier tel que $P(X < a) \leq 0,025$,**

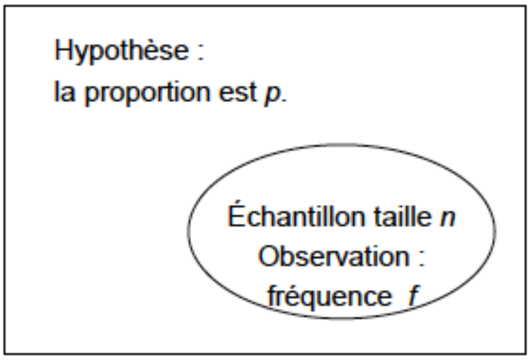
***b* est le plus petit entier tel que $P(X > b) \leq 0,025$.**

ou encore par le système de conditions équivalent :

- *a* est le plus petit entier tel que $P(X \leq a) > 0,025$,
- *b* est le plus petit entier tel que $P(X \leq b) \geq 0,975$.

On remarque que l'intervalle $\left[\frac{a}{n},\frac{b}{n}\right]$ est quasiment centré sur *p* dès que *n* est « assez grand » et que l'intervalle $\left[\frac{a}{n},\frac{b}{n}\right]$ est « quasiment » le même que l'intervalle $\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]$ donné dans le programme de seconde pour les « grandes binomiales » ($n>25$ et $0{,}2<p<0{,}8$ où *n* est la taille de l'échantillon prélevé et *p* est la proportion dans la population du caractère étudié, conditions énoncées dans le programme de seconde).

II- Prise de décision avec la loi binomiale



On considère une population dans laquelle on suppose que la proportion d'un certain caractère est *p*. Pour juger de cette hypothèse, on y prélève, au hasard et avec remise, un échantillon de taille *n* sur lequel on observe une fréquence *f* du caractère.

On rejette l'hypothèse selon laquelle la proportion dans la population est *p* lorsque la fréquence *f* observée est trop éloignée de *p*, dans un sens ou dans l'autre. On choisit de fixer le **seuil à 95 %** de sorte que la probabilité de rejeter l'hypothèse, alors qu'elle est vraie, soit inférieure à 5 %.

Lorsque la proportion dans la population vaut *p*, la variable aléatoire *X* correspondant au nombre de fois où le caractère est observé dans un échantillon aléatoire de taille *n*, suit la loi binomiale de paramètres *n* et *p*.

La **règle de décision** adoptée est la suivante :

si la fréquence observée *f* appartient à l'intervalle de fluctuation au seuil de 95 % $\left[\frac{a}{n},\frac{b}{n}\right]$, on considère que l'hypothèse selon laquelle la proportion est *p* dans la population n'est pas remise en question et on l'accepte ;

sinon, on rejette l'hypothèse selon laquelle cette proportion vaut *p*.

Cette prise de décision repose sur le raisonnement suivant : si la proportion vaut *p* on a, en gros, au moins 95 % de chances que le prélèvement d'un échantillon de taille *n* conduise à une fréquence *f* du caractère dans cet échantillon située dans $\left[\frac{a}{n},\frac{b}{n}\right]$. On sait bien que dans ce cas, compte tenu du hasard, la fréquence réellement observée *f* n'est pas nécessairement égale à *p*, mais qu'elle **fluctue** dans un voisinage de *p*, appelé justement **intervalle de fluctuation**. Un intervalle de fluctuation est donc un intervalle où l'on « s'attend » à trouver la fréquence observée *f*, si l'hypothèse que la proportion est *p* est la bonne.

En conséquence, si la proportion vaut *p*, il y a très peu de chances (environ au plus 5% des échantillons) que cette fréquence observée *f* soit hors de l'intervalle de fluctuation $\left[\frac{a}{n},\frac{b}{n}\right]$. Donc si elle est à l'extérieur de l'intervalle $\left[\frac{a}{n},\frac{b}{n}\right]$, il est cohérent de penser que ce n'est plus le seul fait du hasard cette fois-ci, mais que c'est bien plutôt le signe que l'hypothèse que la proportion est *p* n'est pas la bonne.

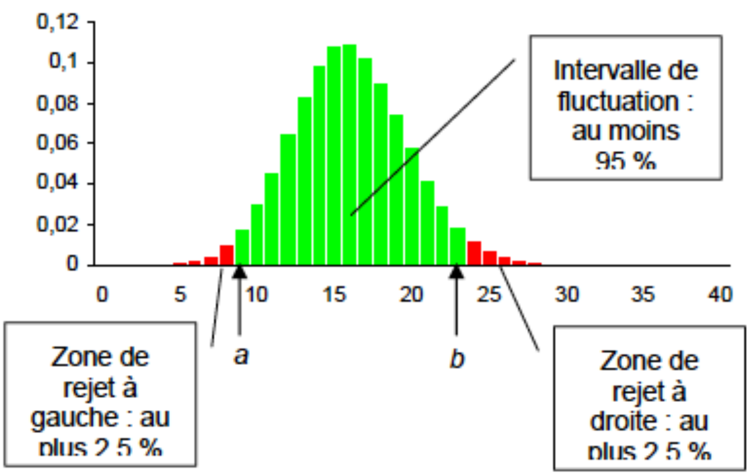
III- Un exemple

Considérons l'exemple suivant. Un médecin de santé publique veut savoir si, dans sa région, le pourcentage d'habitants atteints d'hypertension artérielle est égal à la valeur de 16 % récemment publiée pour des populations semblables. En notant *p* la proportion d'hypertendus dans la population de sa région, le médecin formule l'hypothèse $p = 0{,}16$. Pour vérifier cette hypothèse, le médecin constituera un échantillon de $n = 100$ habitants de la région ; il déterminera la fréquence f d'hypertendus (l'échantillon est prélevé au hasard et la population est suffisamment importante pour considérer qu'il s'agit de tirages avec remise).

Lorsque la proportion dans la population vaut $p = 0{,}16$, la variable aléatoire X correspondant au nombre d'hypertendus observé dans un échantillon aléatoire de taille $n = 100$, suit la loi binomiale de paramètres $n = 100$ et $p = 0{,}16$.

On cherche à partager l'intervalle $[0;100]$, où X prend ses valeurs, en trois intervalles $[0, a-1]$, $[a, b]$ et $[b+1, 100]$ de sorte que la variable aléatoire X prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur. On recherche donc le plus grand entier *a* tel que $P(X < a) \leq 0{,}025$ et le plus petit entier *b* tel que $P(X > b) \leq 0{,}025$.

X suit la loi $\mathcal{B}(100\,;0{,}16)$



D'un point de vue **algorithmique**, il est plus efficace de travailler avec les probabilités cumulées croissantes, que la calculatrice ou le tableur fournissent facilement. En tabulant les probabilités cumulées $P(X \leq k)$, pour *k* allant de 0 à 100, il suffit de déterminer le plus petit entier *a* tel que $P(X \leq a) > 0{,}025$ et le plus petit entier *b* tel que $P(X \leq b) \geq 0{,}975$.

Le calcul, à l'aide de la loi binomiale, de l'intervalle de fluctuation au seuil de 95 %, $\left[\frac{a}{n}, \frac{b}{n}\right]$, de la fréquence des échantillons aléatoires de taille *n*, correspondant à la zone d'acceptation d'une hypothèse sur une proportion, peut ainsi faire l'objet d'une recherche d'algorithme.

C6 =SI(B6>0,025;A6/B$3;"")

	A	B	C	D	E	F	G	H
1	**Intervalle de fluctuation à 95 % d'une fréquence donné par la loi binomiale**							
2	n taille de l'échantillon		p proportion supposée dans la population					
3	**n = 100**		**p = 0,16**					
4								
5	k	P(X<=k)	recherche a	recherche b		**Intervalle de fluctuation à 95 %**		
6	0	2,87873E-08				(selon la loi binomiale)		
7	1	5,37021E-07						
8	2	5,3478E-06				**0,09**	**0,23**	
9	3	3,52815E-05						
10	4	0,000173547						
11	5	0,000679203						
12	6	0,002204197						
13	7	0,006104962						
14	8	0,014742048						
15	9	0,031559427	0,09					
16	10	0,060708551	0,1					
17	11	0,106138316	0,11					
18	12	0,170315459	0,12					
19	13	0,25308401	0,13					
20	14	0,351011275	0,14					
21	15	0,457975907	0,15					
22	16	0,566213927	0,16					
23	17	0,668085005	0,17					
24	18	0,757559073	0,18					
25	19	0,831111691	0,19					
26	20	0,887852282	0,2					
27	21	0,929024593	0,21					
28	22	0,95718574	0,22					
29	23	0,975376792	0,23	0,23				
30	24	0,986493546	0,24	0,24				

On peut, par exemple, procéder sur un tableur comme le montre l'image d'écran.

La cellule B3 contient la valeur de n, taille de l'échantillon. La cellule D3 contient la valeur de p, proportion supposée dans la population.

On a entré en B6 la formule `=SI(A6<=B$3;LOI.BINOMIALE(A6;B$3;D$3;VRAI);"")`

pour tabuler les probabilités P(X □ k) lorsque X suit la loi binomiale de paramètres n et p.

On a entré en C6 la formule `=SI(B6>0,025;A6/B$3;"")`

pour afficher les valeurs de k telles que P(X □ k) dépasse strictement 0,025.

On a entré en D6 la formule `=SI(B6>=0,975;A6/B$3;"")`

pour afficher les valeurs de k telles que P(X □ k) égale ou dépasse 0,975.

Ces trois formules ont été ici recopiées vers le bas jusqu'à la ligne 1 006 (l'algorithme fonctionne pour une valeur maximale de n égale à 1 000, mais on peut, en cas de besoin, recopier plus bas).

L'intervalle de fluctuation au seuil de 95 % est affiché en cellules F8 et G8 contenant les formules `=MIN(C6:C1006) et =MIN(D6:D1006)`.

Dans le cas de l'exemple choisi, on a n = 100 et p = 0,16. L'algorithme fournit $\frac{a}{n} = 0{,}09$ et $\frac{b}{n} = 0{,}23$.

La règle de décision, pour le médecin, sera la suivante :

- si la fréquence observée *f* appartient à l'intervalle de fluctuation [0,09 ; 0,23], on considère que l'hypothèse selon laquelle la proportion d'hypertendus dans la population est *p* = 0,16 n'est pas remise en question et on l'accepte ;
- sinon, on rejette l'hypothèse selon laquelle cette proportion vaut *p* = 0,16.

Exercice :

Monsieur *Z*, chef du gouvernement d'un pays lointain, affirme que 52 % des électeurs lui font confiance. On interroge 100 électeurs au hasard (la population est suffisamment grande pour considérer qu'il s'agit de tirages avec remise) et on souhaite savoir à partir de quelles fréquences, au seuil de 95 %, on peut mettre en doute le pourcentage annoncé par Monsieur *Z*, dans un sens, ou dans l'autre.

1. On fait l'hypothèse que Monsieur *Z* dit vrai et que la proportion des électeurs qui lui font confiance dans la population est 0,52. Montrer que la variable aléatoire *X*, correspondant au nombre d'électeurs lui faisant confiance dans un échantillon de 100 électeurs, suit la loi binomiale de paramètres $n = 100$ et $p = 0{,}52$.

2. On donne ci-contre un extrait de la table des probabilités cumulées $P(X \leq k)$ où X suit la loi binomiale de paramètres $n = 100$ et $p = 0{,}52$.

k	$P(X \leq k) \approx$
40	0,0106
41	0,0177
42	0,0286
43	0,0444
...	...
61	0,9719

a. Déterminer *a* et *b* tels que :

- *a* est le plus petit entier tel que $P(X \leq a) > 0{,}025$;
- *b* est le plus petit entier tel que $P(X \leq b) \geq 0{,}975$.

b. Comparer l'intervalle de fluctuation au seuil de 95 %, $\left[\frac{a}{n}, \frac{b}{n}\right]$, ainsi obtenu grâce à la loi binomiale, avec l'intervalle $\left[p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}}\right]$.

3. Énoncer la règle décision permettant de rejeter ou non l'hypothèse que la proportion des électeurs qui font confiance à Monsieur *Z* dans la population est 0,52, selon la valeur de la fréquence *f* des électeurs favorables à Monsieur *Z* obtenue sur l'échantillon.

4. Sur les 100 électeurs interrogés au hasard, 43 déclarent avoir confiance en Monsieur *Z*. Peut-on considérer, au seuil de 95 %, l'affirmation de Monsieur *Z* comme exacte ?