# 1ère S — Loi binomiale et échantillonnage

**Le mardi 29 août 2017**

Les données sont conformes aux résultats que l'on obtiendrait par le calcul avec le modèle construit en prenant cette valeur du paramètre *p* (cf. test de conformité ou contrôle de qualité dans les entreprises).

## Présentation du chapitre

• Le cours fait le lien entre deux chapitres : les probabilités avec la loi binomiale et les statistiques.

• Il reprend les premières notions d'échantillonnage vues en seconde (notions d'intervalles de fluctuation d'une fréquence, prise de décision…).

• Le cours s'articule en deux parties.

**Partie A :** Loi binomiale et intervalle de fluctuation d'échantillonnage

**Partie B :** Application à la prise de décision sur un échantillon (prise de décision à partir d'une fréquence)

• On étudiera en exercices différentes situations mettant en jeu la prise de décision à partir de la fréquence dans un échantillon.
En arrière-plan, on s'attachera à illustrer les enjeux (économiques, sociaux, sanitaires, politiques,…) de la situation étudiée et de la signification réelle des risques encourus.

## Introduction à l'échantillonnage

Ce chapitre fait suite aux chapitres de probabilités étudiés précédemment.

Si une population est trop importante pour l'étudier complètement, on utilise un **échantillonnage** de cette population.

Si une expérience aléatoire est répétée *n* fois dans des conditions identiques indépendantes, la série statistique obtenue est appelée un **échantillon de taille *n***.
On peut alors déterminer la **distribution des fréquences** de cet échantillon.

Pour une même expérience aléatoire, les distributions de fréquences de deux échantillons de même taille sont le plus souvent différentes. On appelle ceci la **fluctuation d'échantillonnage**.

Le même phénomène s'observe avec le sondage de deux groupes de même taille, mais composés de personnes différentes : les résultats seront le plus souvent différents.

On admettra aussi que plus la taille de l'échantillon est grande, plus la distribution des fréquences s'approche de la distribution « théorique » des fréquences.
Le mot « théorique » se rapporte aux probabilités.
Si l'on considère par exemple le lancer d'une pièce non truquée, lorsque le nombre de lancers augmente, la distribution des fréquences de « pile » et de « face » se rapproche de la distribution théorique idéale $(\frac{1}{2} ; \frac{1}{2})$ (les fréquences ne sont pas forcément égales).



Historiquement, la notion de fluctuation d'échantillonnage est présente dans les travaux de Bernoulli.

**Introduction (rappel de deux exemples étudiés en 2e) :**

① On considère l'épreuve aléatoire qui consiste à lancer 100 fois une pièce non truquée.
On s'intéresse à la fréquence d'apparition de pile.
Cette fréquence ne sera pas égale à 0,5. Mais elle va fluctuer dans un intervalle autour de 0,5.

Plus précisément, si on renouvelle l'expérience aléatoire plusieurs fois, on va obtenir une distribution relativement symétrique des fréquences de piles autour de 0,5.

On constatera alors qu'environ 95 % d'entre elles appartiendront à l'intervalle [0,4 ; 0,6] (centré en 0,5) qui est l'intervalle de fluctuation au seuil approximatif de 95% de la fréquence de pile dans un échantillon de taille 100.

② Idem lancer d'un dé cubique non truqué. On s'intéresse à la fréquence de 1.
On obtient une distribution à peu près symétrique autour de $\frac{1}{6}$, appartient à un intervalle compris autour de $\frac{1}{6}$.

Ces deux exemples montrent une application des intervalles de fluctuation d'une fréquence.

Le mot fluctuation provient du verbe fluctuer (latin *fluctuare*, « être ballotté sur les flots, flotter », que l'on retrouve dans la devise latine de Paris « Fluctuat nec mergitur » qui signifie « Il est battu par les flots, mais ne sombre pas »).

« Le prix de l'essence fluctue au cours du temps. ».

« Mon moral fluctue suivant le temps qu'il fait. ».

# Partie A

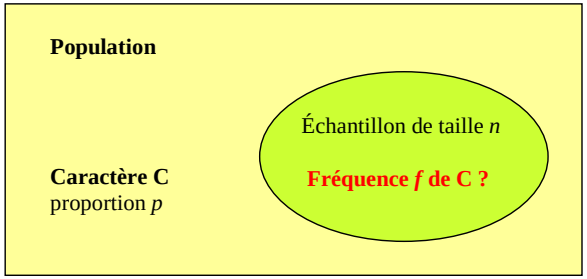
# Loi binomiale et intervalle de fluctuation d'échantillonnage d'une fréquence

## I. Rappel de seconde et utilisation du modèle binomiale

### 1°) Situation étudiée

Dans une population donnée où la proportion d'individus présentant le caractère C est *p*, on prélève un échantillon (au sens aléatoire) de taille *n* (c'est-à-dire que l'on effectue *n* tirages avec remise d'un individu à chaque fois dans la population).

Que peut-on dire de la fréquence *f* de C, sur cet échantillon ?




On peut noter que la proportion d'individus (dans toute la population) présentant le caractère C est le quotient du nombre d'individus présentant le caractère C sur l'effectif total. Cette proportion est souvent exprimée en pourcentage.

On rappelle la définition (fondamentale dans le chapitre) :

$$\textbf{fréquence} = \frac{\textbf{effectif du caractère}}{\textbf{effectif total}}$$

### 2°) Résultat admis en seconde

En classe de seconde, on a observé que, sur un grand nombre d'échantillons de taille *n* (simulés ou non), 95 % au moins fournissent une fréquence *f* appartenant à l'intervalle $\left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$, sous certaines conditions sur *n* et *p*.

Traduit en termes de probabilités, on dispose alors du résultat suivant :

**Pour $n \geqslant 25$ et $0,2 \leqslant p \leqslant 0,8$, lorsqu'on prélève au hasard un échantillon de taille *n* dans une population où la proportion d'un caractère est *p*, la fréquence *f* du caractère sur cet échantillon appartient à l'intervalle $\left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ avec une probabilité supérieure ou égale à 0,95.**



On peut exprimer cette propriété de la manière suivante :

« Il y a au moins 95 % de chances que l'intervalle $\left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ contienne la fréquence du caractère C dans l'échantillon. »

### 3°) Lien avec le schéma de Bernoulli

En classe de première, le tirage au hasard dans la population d'un individu qui peut présenter le caractère C avec une probabilité *p* est une épreuve de Bernoulli de paramètre *p*, où le succès S est l'issue : « avoir C ».

Ainsi, nous pouvons utiliser le modèle binomial pour modéliser la constitution d'échantillons aléatoires.

Le prélèvement au hasard d'un échantillon de taille *n* dans cette population s'assimile à un schéma de Bernoulli de paramètres *n* et *p*, et la variable aléatoire X, qui compte le nombre de succès, c'est-à-dire le nombre d'individus présentant le caractère C, suit la loi binomiale $\mathcal{B}(n ; p)$.

La variable aléatoire $F = \frac{X}{n}$ représente alors la fréquence aléatoire du succès S sur un échantillon de taille *n*.

D'après le résultat de seconde, on a $P\left( F \in \left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right] \right) \geqslant 0,95$ et on dit que $\left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ est un intervalle de fluctuation de F au seuil de 95 %.

### 4°) Prélèvement d'échantillon

Le prélèvement d'un échantillon de taille *n* consiste en *n* répétitions d'un tirage *avec remise* (définition d'un échantillon aléatoire).
Mais si les tirages s'effectuent *sans remise*, l'assimilation à un schéma de Bernoulli est encore possible lorsque la population est suffisamment grande en regard de *n* (taille de l'échantillon).

Autrement dit, dans une grande population qu'on tire avec ou sans remise ne change pas grand-chose (car on a peu de chance de retomber sur le même). C'est évidemment faux pour une petite population.

## II. Définition d'un intervalle de fluctuation à l'aide de la loi binomiale

### 1°) Définition générale d'un intervalle de fluctuation de F

On considère une variable aléatoire X qui suit une loi binomiale $\mathcal{B}(n ; p)$.
On pose $F = \frac{X}{n}$.

F est la variable aléatoire qui représente la fréquence aléatoire du succès.

**Définition générale d'un intervalle de fluctuation de F au seuil de 95 %**

Un **intervalle de fluctuation** de F au seuil de 95 % est un intervalle :
- de la forme $\left[ \frac{a}{n} ; \frac{b}{n} \right]$ où *a* et *b* sont des entiers compris entre 0 et *n* ;
- tel que $P\left( \frac{a}{n} \leqslant F \leqslant \frac{b}{n} \right) \geqslant 0,95$, ce qui équivaut à $P(a \leqslant X \leqslant b) \geqslant 0,95$.

## 2°) Construction d'un intervalle de fluctuation au seuil de 95 %

En pratique, on s'efforce d'obtenir l'intervalle $\left[\frac{a}{n};\frac{b}{n}\right]$ de plus faible amplitude. Pour cela, il suffit de chercher les plus petits entiers naturels $a$ et $b$ tels que $P(X \leqslant a) > 0,025$ et $P(X \leqslant b) \geqslant 0,975$.

On notera que l'on obtient alors un intervalle de fluctuation, à environ 95 %, de la fréquence.
Il s'agit d'un intervalle de fluctuation à au moins 95 % de la fréquence.

La variable aléatoire X compte le nombre d'individus de l'échantillon qui présentent le caractère étudié.
Elle suit la loi binomiale de paramètres $n$ et $p$.
Elle prend les valeurs de tous les entiers naturels de l'intervalle $[0\,;n]$.

On partage l'intervalle $[0\,;n]$ en trois intervalles $[0\,;a-1]$, $[a\,;b]$, $[b+1\,;n]$ de façon que X prenne ses valeurs dans chacun des deux intervalles extrêmes avec une probabilité proche de 0,025 sans jamais la dépasser.

## 3°) Méthode de détermination

X est une variable aléatoire qui suit une loi binomiale $\mathcal{B}(n\,;p)$.

On pose $F = \frac{X}{n}$.
(La variable F ne suit pas la loi binomiale).

Un **intervalle de fluctuation au seuil (approximatif) de 95 % de F** est l'intervalle $\left[\frac{a}{n};\frac{b}{n}\right]$ où :

- $a$ est **le plus petit entier naturel** tel que $P(X \leqslant a) > 0,025$ ;
- $b$ est **le plus petit entier naturel** tel que $P(X \leqslant b) \geqslant 0,975$.

Il s'agit d'un seuil approximatif de 95 %.

La probabilité de l'événement $(a \leqslant X \leqslant b)$ (qui est aussi égale à la probabilité de l'événement $\left(\frac{a}{n} \leqslant F \leqslant \frac{b}{n}\right)$) est supérieure ou égale à 0,95 (cf. démonstration).
En pratique, avec la méthode de détermination donnée ici, la probabilité de l'événement $(a \leqslant X \leqslant b)$ n'est pas forcément égale à 0,95 mais est, en tous cas, proche de 0,95.

L'existence des entiers naturels $a$ et $b$ provient de la croissance de la fonction de répartition F de la variable aléatoire X (par définition, $F : x \mapsto P(X \leqslant x)$.

Pour trouver $a$ et $b$, on peut utiliser un tableur ou la calculatrice.

On notera qu'il n'y pas de formule pour détermine l'intervalle de fluctuation au seuil de fluctuation.



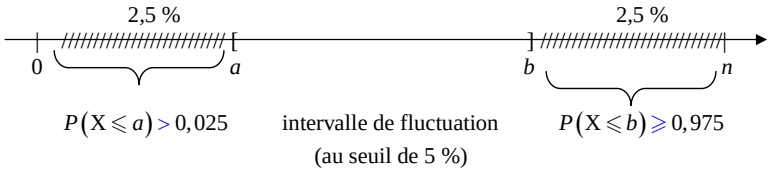
## 3°) Démonstrations

• On déduit directement de la définition que $P(X \leqslant a-1) \leqslant 0,025$ et $P(X \geqslant b+1) \leqslant 0,025$.

• On a : $P(a \leqslant X \leqslant b) = P(X \leqslant b) - P(X \leqslant a-1)$.
Or $-P(X \leqslant a-1) \leqslant -0,025$ et $P(X \leqslant b) - P(X \leqslant a-1) \geqslant 0,975 - 0,025$ soit $P(a \leqslant X \leqslant b) \geqslant 0,95$.

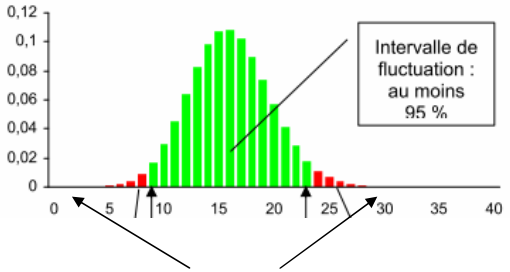
## 4°) Commentaires

• L'intervalle de fluctuation à 95 % est un intervalle qui contient au moins 95 % des fréquences observées dans les échantillons de taille $n$.
Ceci signifie qu'il y a un **risque de 5 %** pour cette fréquence de ne pas se trouver dans cet intervalle.




• L'intérêt du cours de première est de donner un **intervalle de fluctuation « exact »** (par opposition à l'intervalle de fluctuation donné en seconde qui est un intervalle « approximatif ») sans condition sur les valeurs de $n$ et de $p$ (valables en particulier sur de « petits » échantillons).
Le lien avec l'intervalle de fluctuation donné en seconde sera vu en Terminale (notion d'intervalle de fluctuation « asymptotique »).

• Il est possible de visualiser l'intervalle de fluctuation sur le diagramme en bâtons de la loi binomiale (cf. **partie B**) comme le montre le graphique ci-dessous pour une variable aléatoire X qui suit la loi binomiale $\mathcal{B}(100\,;0,16)$.




En fait, ça fait une valeur (mais très proche de 0).

On peut visualiser aisément le principe de cette méthode sur ce graphique.
On a délaissé un peu moins de 2,5 % des valeurs « d'un côté et de l'autre » (en rouge sur le graphique du haut) pour en conserver au moins 95 % dans l'intervalle restant.

• Certains sites Internet permettent d'obtenir directement l'intervalle de fluctuation en rentrant les paramètres de la loi binomiale (c'est le cas par exemple du site Euler de l'académie de Versailles).

## III. Exemple de détermination pratique d'un intervalle de fluctuation d'échantillonnage

**Exercice-type :**

Dans une urne contenant 3 boules blanches et 7 boules noires, on effectue 100 tirages au hasard avec remise (« urne de Bernoulli »).

On veut déterminer un intervalle de fluctuation au seuil de 95 % de la fréquence d'une boule blanche dans l'échantillon prélevé grâce à la loi binomiale.

Le nombre X de boules blanches suit la loi binomiale $\mathcal{B}(100 ; 0{,}3)$.

La fréquence de « boule blanche » est donnée par la variable aléatoire $F = \frac{X}{100}$.

On cherche :

- le plus petit entier naturel $a$ tel que $P(X \leqslant a) > 0{,}025$ ;
- le plus petit entier naturel $b$ tel que $P(X \leqslant b) \geqslant 0{,}975$.

On utilise les valeurs des probabilités cumulées de X (ou fonction de répartition de X) obtenues grâce à la calculatrice ou à un tableur.



| $k$ | $P(X \leqslant k)$ |
|---|---|
| **16** | 0,000968865 |
| **17** | 0,0022162933 |
| **18** | 0,004522639 |
| **19** | 0,008887208 |
| **20** | 0,016462853 |
| **21** | **0,028831253** |
| **22** | **0,047865739** |
| **23** | **0,075530767** |
| **24** | **0,113570182** |
| **25** | **0,163130104** |
| **26** | **0,22439924** |
| **27** | **0,296366161** |
| **28** | **0,376778179** |
| **29** | **0,462339736** |
| **30** | **0,549123601** |
| **31** | **0,633107986** |
| **32** | **0,710718556** |
| **33** | **0,779257761** |
| **34** | **0,83714171** |
| **35** | **0,883921394** |
| **36** | **0,920119958** |
| **37** | **0,946954414** |
| **38** | **0,966021002** |
| **39** | **0,979011424** |
| **40** | 0,987501593 |
| **41** | 0,992826437 |
| **42** | 0,996032211 |
| **43** | 0,997885383 |
| **44** | 0,998914254 |

**Pour obtenir une telle table sur calculatrice TI,** pour une loi binomiale $\mathcal{B}(n\,;p)$, il faut rentrer dans Y1 : binomFrép (dans 2nde var (distrib)) puis mettre *n*, *p*, X et après, aller dans « table ». Voir fin du chapitre.

On obtient $a=21$ et $b=39$.

On rappelle que $n=100$.

$\frac{a}{n}=\frac{21}{100}=0,21$ et $\frac{b}{n}=\frac{39}{100}=0,39$.

Donc $[0,21\,;0,39]$ est un intervalle de fluctuation au seuil de 95 % de la fréquence de « boule blanche ».

**Interprétation :**

Le risque de voir la fréquence de « boule blanche » sur un échantillon de sortir de cet intervalle est donc inférieur à 5 %.

Il serait intéressant de visualiser cet intervalle de fluctuation sur le diagramme en bâtons de la loi binomiale $\mathcal{B}(100\,;0,3)$.

On constate que la probabilité de tirer une boule blanche $p=0,3$ appartient bien à l'intervalle [0,21 ; 0,39]. De plus, on constate que l'intervalle [0,21 ; 0,39] est centré en 0,3. Ce n'est pas le cas en général pour les intervalles de fluctuation déterminés à l'aide de la loi binomiale.
Dans notre cas, nous obtenons un intervalle de fluctuation centré en 0,3 d'amplitude 0,18.

**Commentaire :**

D'un point de vue algorithmique, les valeurs de *a* et *b* peuvent être trouvées grâce à un programme en utilisant les probabilités cumulées (détermination de « valeurs seuils »).

## IV. Autres seuils

Il est possible de définir l'intervalle de fluctuation d'une fréquence à un autre seuil que 95 %.

• Au lieu du coefficient 95 %, on peut choisir d'autres coefficients.
Le plus fréquemment utilisé après 95 % est 99 %, soit un seuil de risque de 1 %. On partage alors l'ensemble des valeurs prises par X en trois parties :

- deux « zones de rejet » A et C telles que $P(X\in A)\leqslant 0,005$ et $P(X\in C)\leqslant 0,005$

et

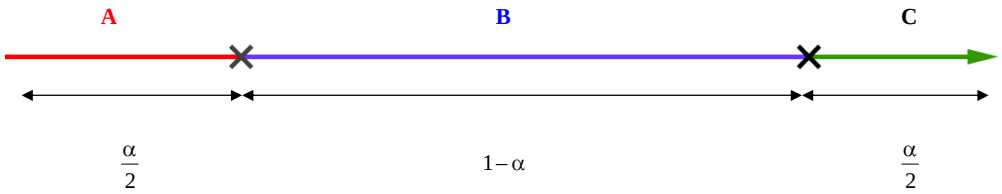
- une « zone centrale » B telle que $P(X\in B)\geqslant 0,99$.

Comme précédemment, les réels *a* et *b* de la propriété vérifient : *a* est le plus petit entier tel que $P(X\leqslant a)>0,005$ et *b* est le plus petit entier tel que $P(X\leqslant b)>0,995$.



• **Cas général :** on choisit un seuil de risque $\alpha$.
Alors, en reprenant les calculs précédents, les probabilités $P(X\in A)$ et $P(X\in C)$ doivent être inférieures ou égales à $\frac{\alpha}{2}$ afin que $P(X\in B)$ soit au moins égale à $1-\alpha$.

Ainsi *a* et *b* seront les plus petits entiers tels que $P(X\leqslant a)>\frac{\alpha}{2}$ et $P(X\leqslant b)\geqslant 1-\frac{\alpha}{2}$.




On exige souvent en général un seuil de fluctuation élevé (en général 95 % ou plus).

## V. Comparaison de l'intervalle de fluctuation obtenu avec la formule de seconde et l'intervalle de fluctuation obtenu avec la loi binomiale

**Objectif :** déterminer un intervalle de fluctuation à l'aide de la loi binomiale et comparer avec celui donné en seconde

**Exemple :** pour $p=0,25$ et $n=100$.

• L'intervalle obtenu avec le résultat de seconde est l'intervalle centré en *p* : $I=\left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$.

On calcule les bornes de I :

$p-\frac{1}{\sqrt{n}}=0,25-\frac{1}{\sqrt{100}}=0,15$

$p+\frac{1}{\sqrt{n}}=0,25+\frac{1}{\sqrt{100}}=0,35$

Donc $I=[0,15\,;0,35]$.

• Déterminons maintenant avec la loi binomiale $\mathcal{B}(100\,;0,25)$ un intervalle de fluctuation de F au seuil de 95 % (avec la définition qui a été donné).

On utilise un tableau donnant la loi de probabilité de X grâce à la calculatrice ou à un tableur.

| $k$ | $P(X \leqslant k)$ |
|---|---|
| … | … |
| 15 | 0,0111 |
| 16 | 0,0211 |
| **17** | **0,0376** |
| 18 | 0,0630 |
| … | … |
| 33 | 0,9724 |
| **34** | **0,9836** |
| 35 | 0,9906 |
| **36** | **0,9948** |

Dans le tableau précédent, on lit que :

- le plus petit entier $a$ tel que $P(X \leqslant a) > 0,025$ est $a = 17$ ;

- le plus petit entier $b$ tel que $P(X \leqslant b) \geqslant 0,975$ est $b = 34$.

L'intervalle de fluctuation cherché est $J = [0,17 ; 0,34]$.

• **Comparaison :**

- Les deux démarches produisent des intervalles très voisins.

- On remarque que l'intervalle J est inclus dans l'intervalle I.

- J étant un intervalle de fluctuation de F au seuil de 95 % (les valeurs de F se situent dans J avec une probabilité supérieure ou égale à 0,95), cela est encore vrai pour l'intervalle I qui le contient. Ceci conforte le résultat vu en classe de seconde, plus simple à utiliser que celui issu de la loi binomiale mais plus approximatif.

- L'intervalle I est centré en $p$ ; ce n'est pas le cas de J, sur cet exemple. Mais il peut arriver que J soit lui aussi centré.

Plus généralement, d'autres comparaisons sont possibles à partir du tableau suivant, qui donnent les bornes de I et J, pour $n = 1000$ et $p$ variant de 0,1 à 0,9.



| $p$ | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 |
|---|---|---|---|---|---|---|---|---|---|
| $\frac{a}{n}$ | 0,082 | 0,176 | 0,272 | 0,370 | 0,469 | 0,570 | 0,671 | 0,775 | 0,881 |
| $\frac{b}{n}$ | 0,119 | 0,225 | 0,329 | 0,430 | 0,531 | 0,630 | 0,728 | 0,824 | 0,918 |
| $p - \frac{1}{\sqrt{n}}$ | 0,068 | 0,168 | 0,268 | 0,368 | 0,468 | 0,568 | 0,668 | 0,768 | 0,868 |
| $p + \frac{1}{\sqrt{n}}$ | 0,132 | 0,232 | 0,332 | 0,432 | 0,532 | 0,632 | 0,732 | 0,832 | 0,932 |

L'intérêt de $\left[\frac{a}{n} ; \frac{b}{n}\right]$, calculé à partir de la loi binomiale, est de fournir un intervalle convenable **pour toutes les valeurs de *n* et de *p***, alors que l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$ **n'est pas adapté** pour les « petites binomiales ».

L'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$, facilement calculable, résulte d'approximations, alors que la loi binomiale est la loi exacte correspondant à la situation.

# Partie B

# Application à la prise de décision sur un échantillon

## I. Présentation du problème

### 1°) Situation générale

Dans une population, **on suppose** qu'un caractère est présent dans la proportion $p$.

Pour juger de cette hypothèse (le terme « hypothèse » est pris ici dans son sens habituel dans le cadre des statistiques et non dans son sens usuel en mathématiques), on prélève au hasard et avec remise, un échantillon de taille $n$ et on observe que la fréquence du caractère étudié est $f$.

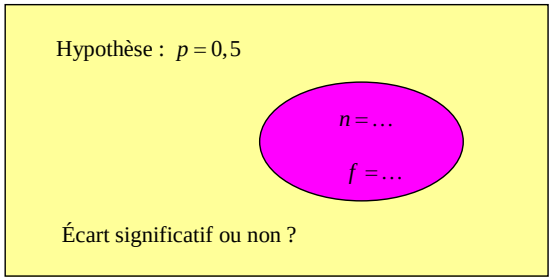
**Population**
Hypothèse : proportion $p$ d'un caractère

**Échantillon**
Taille : $n$
Fréquence $f$ du caractère

On cherche à savoir pour quelles valeurs de $f$ on pourra considérer qu'elles sont suffisamment éloignées de $p$ pour rejeter l'hypothèse.

### 2°) Exemple de prise de décision à partir d'un échantillon

On cherche à savoir si une pièce est équilibrée.




- On fait l'« hypothèse » (au sens des statistiques) que la pièce est équilibrée, donc que la probabilité d'obtenir PILE est $p = 0,5$.

- On lance $n$ fois cette pièce et on détermine la fréquence de PILE sur l'échantillon obtenu.

- On se fixe un seuil, par exemple 95 %, et on détermine l'intervalle de fluctuation I au seuil de 95 % à l'aide de la loi $\mathcal{B}(n\,;p)$.

- On prend une décision :

  Si $f$ n'est pas dans I, on rejette l'hypothèse de pièce équilibrée avec un risque de se tromper dans 5 % des cas.

  Si $f$ appartient à I, on ne rejette pas l'hypothèse de pièce équilibrée (on évite de dire qu'on « l'accepte », car le risque de se tromper en l'acceptant est inconnu).

## II. Règle de décision

### 1°) Énoncé (pour un risque d'erreur de 5 %)

On reprend les notations du **I. 1°)**.

• Si $f \in \left[\frac{a}{n}\,;\frac{b}{n}\right]$, alors **on ne peut pas rejeter l'hypothèse** selon laquelle la proportion du caractère dans la population est égale à $p$ c'est-à-dire que les données observées sont déclarées *compatibles* avec le modèle construit en prenant cette valeur pour paramètre.

• Si $f \notin \left[\frac{a}{n}\,;\frac{b}{n}\right]$, alors **on rejette cette hypothèse** au risque d'erreur de 5 % c'est-à-dire que les données observées sont déclarées *incompatibles* avec le modèle construit en prenant cette valeur pour paramètre.



### 2°) Raisonnement sous-jacent

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 l'intervalle de fluctuation $\left[\frac{a}{n}\,;\frac{b}{n}\right]$.

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.
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.

### 3°) Commentaires

• « Au risque d'erreur de 5 % » signifie que la probabilité de rejeter l'hypothèse alors qu'elle est vraie est strictement inférieure à 5 %.

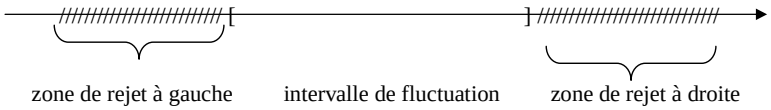
• La règle de décision doit être établie avant l'observation de l'échantillon.

• La règle de décision peut être énoncée au seuil de risque 5 %, mais aussi 1 %, ou bien pour tout autre seuil de risque.

• Lorsque l'on rejette une hypothèse avec un risque de se tromper de 5 %, on ne dit pas qu'il y a une probabilité égale à 0,05 de se tromper.
En effet, il n'y a rien d'aléatoire dans la prise de décision.

• La méthode est basée sur la comparaison empirique-théorique : le résultat d'une expérience (fréquence mesurée, empirique) est comparé au résultat obtenu à partir d'un modèle (loi de probabilité, théorique).
La modèle ici est celui de la loi binomiale.
Ce type de raisonnement est appelé « preuve statistique ».

• La méthode exposée dans le cours est une méthode de **« test bilatéral »**, liée à la forme du diagramme en bâtons de la loi binomiale pour les valeurs des paramètres liées à l'exemple.
On parle alors de « zone de rejet à gauche » et de « zone de rejet à droite ».




Parfois, on utilisera un **« test unilatéral »** ; cela est lié à la forme du diagramme en bâtons de la loi binomiale étudiée (voir exercices).

### 4°) Point histoire

Ce sont les statisticiens Fischer, **Jerzy Neyman** (1894-1981) et **Egon Pearson** (1895-1980) qui ont proposé vers 1930 une démarche de décision universellement admise, à la base de la théorie des tests statistiques.
On peut également ajouter le nom de **Ronald Aylmer Fischer** (1890-1962).

## III. Exercice-type : rejeter ou non une hypothèse

### 1°) Énoncé

Un candidat à une prochaine élection pense que 54 % des électeurs lui apporteront leur voix. Un sondage est réalisé auprès de 100 électeurs (le choix des électeurs est assimilé à un tirage au hasard et avec remise) et on note $f$ la fréquence, dans cet échantillon, des personnes qui voteront pour le candidat.
**On fait l'hypothèse** que la proportion des électeurs qui voteront pour le candidat est effectivement $p = 0,54$.

a) On note X la variable aléatoire qui compte le nombre d'électeurs qui voteront pour ce candidat dans l'échantillon.
Expliquer pourquoi X suit une loi binomiale. Préciser ses paramètres.
b) Déterminer l'intervalle de fluctuation au seuil de 95 % de la fréquence $f$ observée.
c) Énoncer la règle de décision permettant de rejeter ou non l'hypothèse $p = 0,54$, selon la valeur de la fréquence $f$ des électeurs favorables au candidat dans l'échantillon.
d) On observe que, sur les 100 électeurs interrogés, 43 déclarent voter pour le candidat.
Considère-t-on alors l'affirmation du candidat comme exacte ou non ?

### 2°) Solution

a) L'épreuve de Bernoulli consiste à demander son opinion sur le candidat à un électeur.
Le succès est S : « l'électeur votera pour le candidat » et $P(\mathrm{S}) = 0,54$.

Le schéma de Bernoulli consiste à répéter 100 fois cette épreuve dans des conditions d'indépendance (tirage avec remise).
Donc X suit la loi binomiale de paramètres $n = 100$ et $p = 0,54$.

b) Voici ci-dessous un extrait de la table des probabilités $P(\mathrm{X} \leqslant k)$ pour les valeurs $k$ prises par X.

| $k$ | $P(\mathrm{X} \leqslant k)$ |
|---|---|
| … | … |
| 42 | 0,010607189 |
| 43 | 0,017706913 |
| **44** | **0,028503825** |
| … | … |
| 62 | 0,956686279 |
| 63 | 0,972421632 |
| **64** | **0,983100721** |
| … | … |



On lit dans cette table :

• le plus petit entier $a$ tel que $P(\mathrm{X} \leqslant a) > 0,025$ ; on obtient $a = 44$ ;
• le plus petit entier $b$ tel que $P(\mathrm{X} \leqslant b) \geqslant 0,975$ ; on obtient $b = 64$.

L'intervalle de fluctuation au seuil de 95 % de $f$ est donc l'intervalle $[0,44 ; 0,64]$.

c) Si la fréquence $f$ appartient à l'intervalle $[0,44 ; 0,64]$, alors on ne peut pas rejeter l'hypothèse $p = 0,54$, sinon cette hypothèse est rejetée au risque d'erreur de 5 %.

d) Dans ce cas, $f = 0,43$, $f$ n'appartient pas à l'intervalle $[0,44 ; 0,64]$ et l'hypothèse $p = 0,54$ est rejetée. On considère l'affirmation du candidat inexacte.
Les données observées sont déclarées incompatibles avec le modèle construit en prenant la valeur 0,54 pour paramètre.

**Commentaire :**
On remarque dans cet exemple que l'intervalle de fluctuation $\left[ p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$ au seuil de 95 % donné en classe de seconde coïncide exactement avec l'intervalle $[0,44 ; 0,64]$ obtenu ici avec une loi binomiale.

# Appendice

## L'outil statistique : une aide à la décision

Marc Lavielle, directeur de recherche, INRIA, Saclay

---

### ↘ Le rôle de la statistique dans les études de toxicité

« Nous allons ici nous limiter à une étude particulière dont l'objectif est d'explorer une éventuelle toxicité du maïs OGM. On nourrit pendant 90 jours des groupes de rats avec différents régimes (maïs OGM ou non OGM) et différentes doses (11 % et 33 %). On mesure de très nombreux paramètres […]. On va ensuite regarder si des dissemblances apparaissent entre les groupes tests et les groupes témoins […]. Bien évidemment, des différences sont toujours observées et une partie de ces différences observées est simplement due au hasard (c'est-à-dire à la façon dont ont été constitués les échantillons). Une première question se pose alors :

• *Ces différences ne sont-elles pas dues qu'au hasard, ou bien une partie de ces différences peut-elle être expliquée par la différence de régime (OGM versus non OGM) ?*
Cette question est d'ordre purement statistique. On met en œuvre des tests statistiques de comparaison pour tenter de répondre à cette question. La procédure classique consiste à comparer pour chaque paramètre, les moyennes dans les groupes témoins et tests. On pose alors comme hypothèse […] que les moyennes sont identiques dans les deux groupes. On ne peut ensuite rejeter cette hypothèse que si la différence observée est suffisamment importante. Ce seuil est traditionnellement fixé à une valeur telle que […] la probabilité de conclure à tort à un effet OGM soit de 5 %.

• *Si des différences sont considérées comme statistiquement significatives […], faut-il alors conclure à l'existence d'un risque pour notre santé ?*
Ce n'est plus du tout une question d'ordre statistique, et seul le toxicologue est en mesure d'évaluer si les différences qu'il observe peuvent indiquer ou non des signes de toxicité […]. Le statisticien ne fait que lever des drapeaux orange pour les paramètres sur lesquels il soupçonne un effet OGM ; c'est ensuite au toxicologue de lever le drapeau rouge s'il observe une configuration particulière de drapeaux oranges levés.

### ↘ Le rôle du statisticien dans un débat comme celui des OGM est multiple

• *Rappeler que nous sommes dans un environnement incertain*
Le peu d'information que l'on peut tirer de ces études ne permet pas de conclure que le maïs OGM est dangereux, mais il n'autorise pas pour autant à conclure fermement à son innocuité. La première fonction du statisticien consiste à évaluer le juste niveau des incertitudes.

• *Apporter de bonnes réponses à de bonnes questions*
[…] Le fait de changer de régime provoque inévitablement des modifications de nombreux paramètres physiologiques, et ce, indépendamment du caractère toxique de ces régimes. Même infimes, ces différences existent et il est donc paradoxal d'en tester l'existence ! La vraie question qu'il convient de se poser est de savoir si ces différences sont suffisamment importantes pour être associées à un effet toxique. Le bon outil statistique pour répondre à cette question n'est pas le test de comparaison des moyennes (particulièrement favorable à l'industriel qui part du principe qu'il n'existe pas d'effet OGM) mais le test de bio-équivalence qui protège davantage le consommateur en partant de l'hypothèse qu'il existe un effet OGM préoccupant : c'est alors à l'expérience de démontrer qu'il n'en est rien.



D'autre part, se contenter d'évaluer la probabilité de se tromper en concluant à tort l'existence d'un effet OGM (donc risquer de ne pas commercialiser un aliment sans danger) n'est pas suffisant. Il faut systématiquement évaluer la puissance du test, c'est-à-dire la probabilité de conclure à un effet OGM lorsque cet effet existe réellement (dans le but de ne pas commercialiser un aliment dangereux) […].

• *Rappeler le rôle limité de la statistique*
C'est un outil d'aide à la décision, mais pas un outil de décision ! Ce n'est pas la statistique qui permet de conclure si un OGM est dangereux ou non pour la santé humaine […]. La statistique est là uniquement pour aider le toxicologue à évaluer correctement les risques de se tromper en concluant sur l'absence ou la présence d'effets négatifs.
C'est finalement au gestionnaire du risque de prendre une décision, qui va prendre en compte bien sûr l'évaluation des différents risques, mais également d'autres critères de type économique ou social. »

Source : *Maths à venir Express*, CIJM

Article tiré du livre Math'x 1ère S Programme 2011 (Édition Didier) page 178

# Appendice sur les calculatrices

# Programmes sur calculatrice

Ces programmes permettent de déterminer les valeurs de *a* et de *b* correspondant à l'intervalle de fluctuation au seuil de 95 % trouvé à partir de la loi binomiale.

### Programme fluctuation

| Calculatrice TI | Calculatrice Casio |
|---|---|
| : Prompt N<br>: Prompt P<br>: 0 → I<br>: 0 → J<br>: While binomFrép(N, P, I) ⩽ 0,025<br>: I + 1 → I<br>: End<br>: While binomFrép(N, P, J) < 0,975<br>: J + 1 → J<br>: End<br>: Disp "I =", I<br>: Disp "J =", J | "N" : ? → N ↵<br>"P" : ? → P ↵<br>: 0 → I↵<br>: 0 → J↵<br>: binominalCD(N, P) → List 1↵<br>: While List1[I + 1] ⩽ 0,025↵<br>: I + 1 → I↵<br>: WhileEnd ↵<br>: While List1[J + 1] < 0,0975 ↵<br>: J + 1 → J↵<br>: WhileEnd↵<br>: I◢<br>: J ◢ |



### Memo pour déterminer un intervalle de fluctuation à 95 % à l'aide de la calculatrice (modèle TI)

Dans une urne contenant 3 boules blanches et 7 boules noires, on effectue 100 tirages au hasard avec remise.

Déterminer un intervalle de fluctuation au seuil de 95 % de la fréquence d'une boule blanche dans l'échantillon prélevé grâce à la loi binomiale.

On va utiliser la fonction de répartition de la loi binomiale de paramètres $n = 100$ et $p = 0,3$.

$f(x)$ $Y1 = \left|\begin{array}{l}\text{binomFRép}(100, 0.3, X)\\ \text{binomcdf}(100, 0.3, X)\end{array}\right.$ [aller dans 2nde var (distrib) et choisir B : binomFRép( ]

Modèles de calculatrice plus récents :

xvalue : mettre X
trials : mettre la valeur de *n*
p : mettre la valeur de la probabilité *p*

ou

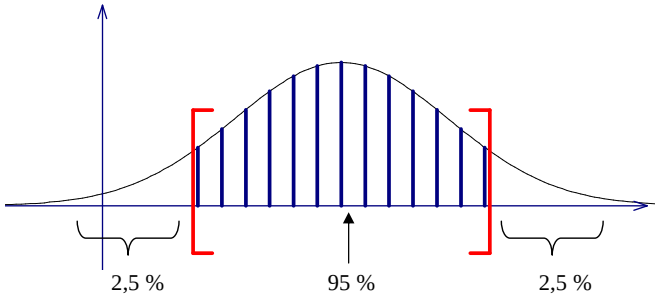
nbreEssais : mettre la valeur de n
p : mettre la valeur de la probabilité *p*
valeur de x : mettre X

On règle la table en partant de 0 avec un pas de 1.

On cherche dans la table la plus petite valeur de X telle que $Y1 > 0,025$ (*a*).

On cherche dans la table la plus petite valeur de X telle que $Y1 \geqslant 0,975$ (*b*).

L'intervalle de fluctuation au seuil de 95 % est $\left[\frac{a}{100}; \frac{b}{100}\right]$ c'est-à-dire $[0,21; 0,39]$.


2,5 %
95 %
2,5 %