Chapitre 10

Echantillonage

Avant toutes choses, il est important d'introduire quelques points de vocabulaire.

10.1 Echantillons et fréquences

Pour avoir une meilleur intuition à propos de la notion **d'échantillon**, voici quelques exemples.

Exemple 10.1.1.
- Lancer 100 fois un dé et noter la liste des résultats obtenus.
- Prélever 100 ampoules d'une chaine de fabrication. Tester, puis noter à chaque fois si elles sont conforme ou non.
- Interroger 100 personnes au hasard et noter à chaque fois leur couleur préférée.
- ...

Dans les trois situations précédentes, nous avons noté les résultats d'une expérience aléatoire qui a été répété $n = 100$ fois de manière indépendante dans les mêmes conditions.

Définition 10.1.1. *Un **échantillon** de taille n est la liste de n résultats obtenues lors de n **répétitions indépendantes** de la même expérience aléatoire.*

Remarque. Lorsque l'expérience aléatoire ne présente que deux issues possibles (obtenir pile ou face, gagner ou perdre, etc ...), nous retrouvons le cas des épreuves de Bernoulli.

En pratique nous faisons face à une liste de valeurs à partir desquelles nous pouvons déterminer la **fréquence d'apparition** des différentes issues de l'expérience.

Exemple 10.1.2. Si nous avons à disposition un échantillon de taille $n = 1000$ contenant les résultats obtenus après avoir effectué une série de 1 000 piles ou faces. Supposons que nous ayons obtenu 531 fois l'issue pile, cela signifie que la **fréquence observée** de l'issue « pile » vaut

$$f = \frac{531}{1000} = 0,531.$$

Ces nouvelles notions soulèvent plusieurs questions auxquelles nous allons répondre dans ce chapitre. Pour cela, reprenons les données de l'exemple précédent. A priori **nous ne savons pas si la pièce est équilibrée.** Autrement dit,

$$\mathbb{P}(\text{pile}) = p \in [0;1] \quad \text{avec} \quad p \quad \text{une quantité } \textbf{inconnue} \text{ que nous aimerions estimer.}$$

Cependant nous savons, d'après *la loi forte des grands nombres* de Kolmogorov, que plus la taille de l'échantillon augmente plus la fréquence observée f se rapproche de la valeur théorique inconnue p. Il est alors naturel de s'interroger :

1. serait-il possible d'utiliser la fréquence f et la taille de l'échantillon n pour proposer un intervalle I_n contenant p avec une marge d'erreur fixée au préalable ? Il s'agit d'un **problème d'estimation** et de **construction d'un intervalle de confiance**.
2. Si jamais la pièce était équilibrée, nous savons que $p = \frac{1}{2}$. Or, dans l'exemple précédent, nous avons obtenu $f = 0,531$; qu'est-il possible d'en déduire ? La pièce est-elle équilibrée ? Ce genre de questionnements intervient dans la **théorie des tests en statistiques** dont nous allons présenter certains aspects dans ce chapitre.

10.2 Estimation et intervalle de confiance

Par la suite, nous supposons avoir à disposition un échantillon de taille n. Poursuivons à présent l'étude amorcée dans l'introduction du chapitre en débutant par la construction d'intervalle de confiance.

Exemple 10.2.1.
1. Toujours avec l'exemple du pile ou face. Nous cherchons à estimer la valeur théorique $\mathbb{P}(\text{pile}) = p \in [0;1]$ (a priori nous ne savons pas si la pièce est équilibrée ou non) à l'aide d'un échantillon et des fréquences observées.
2. Plus généralement, nous pouvons chercher à déterminer la proportion p (inconnue) d'apparition d'un caractère dans un échantillon.

Une réponse acceptable serait de dire : la valeur théorique p **se trouve dans un intervalle** I_n (qui dépendrait de la taille de l'échantillon n et de la fréquence observée f) et cette affirmation se fait avec **une marge d'erreur de** 5%. Le théorème suivant apporte une réponse à cette problématique.

Théorème 35 (Intervalle de confiance)**.** *Dans le contexte précédent, pour tout $n \geq 1$, l'intervalle*

$$I_n = \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right]$$

***contient la valeur théorique** p **avec un risque de** 5% ; f correspond à la fréquence d'apparition observée du caractère p dans l'échantillon. L'intervalle I_n est appelé **intervalle de confiance**.*

Remarque. Implicitement, nous utilisons une loi Binomiale $X \sim B(n;p)$ (associée à l'épreuve de Bernoulli de succès « le caractère étudié apparait ») et $f = \frac{X}{n}$.

Voyons, à l'aide d'exemples, ce que ce théorème permet de dire.

Exemple 10.2.2. Un sac opaque contient un très grand nombre de boules rouges ou bleues indiscernables au toucher. Lors d'un tirage de 100 boules, nous avons obtenues 41 boules rouges et 59 boules bleues. C'est pourquoi, la fréquence observée de boules rouges vaut

$$f = \frac{41}{100} = 0,41.$$

D'après le théorème 35 précédent, nous savons que la véritable proportion p de boules rouges est contenue (avec un niveau de confiance de 95%) dans l'intervalle

$$I_n = \left[0,41 - \frac{1}{\sqrt{100}}; 0,41 + \frac{1}{\sqrt{100}} \right] = [0,40; 0,42].$$

L'exemple précédent et sa conclusion soulèvent de nouvelles questions. Voyons cela sur deux autres exemples.

Exemple 10.2.3. Les 10 000 employés d'une entreprise doivent voter pour élire, parmi deux candidats (notés numéro 1 et numéro 2), un nouveau délégué syndical . Curieuse du résultat, la direction a interrogé un échantillon de 100 employés sur leur choix et 54% ont répondu qu'ils avaient voté pour le candidat numéro 2.

Voyons ce qu'il est possible de déduire de cette information. Ici, $n = 100$ et p est un nombre inconnu (correspondant à la véritable proportion de personnes ayant choisi le candidat numéro 2). La fréquence observée vaut $f = 0,54$ et $\frac{1}{\sqrt{n}} = 0,1$. Ainsi, l'intervalle de fluctuation associé à p est de la forme

$$I_n = [0,44; 0,64]$$

D'après le théorème 35 précédent, il y a donc de très grandes chances (au moins 95%) pour que la proportion réelle p de personnes ayant voté pour le candidat numéro 2 se trouve dans cet intervalle. **Il n'est cependant pas possible de savoir qui va l'emporter** puisque les deux situations peuvent se produire ($p < \frac{1}{2}$ entrainant la victoire du candidat 1 ou $p > \frac{1}{2}$ entrainant la victoire du candidat 2).

Nous aurions donc besoin d'un critère permettant de tester l'hypothèse

$$H_0 \ : \ p < \frac{1}{2} \quad \text{contre} \quad H_1 \ : \ p > \frac{1}{2}.$$

En réfléchissant un peu, le théorème 35 fournissant l'intervalle de confiance permet parfois de répondre à cette question.

Exemple 10.2.4. Supposons que nous souhaitions vérifier qu'un dé soit truqué ou non et que nous ayons à disposition un échantillon de 2 500 lancers. Observons au passage le fait suivant : si le dé n'est pas truqué, il y a autant de chances d'obtenir un nombre pair qu'un nombre impair. Autrement dit :

$$\mathbb{P}(\text{pair}) = p = \frac{1}{2}.$$

Supposons de plus que nous ayons un échantillon de 2 500 lancers dans lequel nous observons 1150 résultats pairs. Autrement dit, la fréquence observée (du nombre de résultats pairs obtenus sur les 2500 lancers) vaux $f = \frac{1150}{2500} = 0,46$. Le théorème 35 nous assure alors que

$$p \in \left[f - \frac{1}{\sqrt{n}}; f - \frac{1}{\sqrt{n}}\right] \quad \Longleftrightarrow \quad f \in \left[p - \frac{1}{\sqrt{n}}; p - \frac{1}{\sqrt{n}}\right]$$

avec une marge d'erreur de 5%. Si la pièce était équilibrée (i.e. $p = \frac{1}{2}$), nous devrions avoir

$$f \in \left[p - \frac{1}{\sqrt{n}}; p - \frac{1}{\sqrt{n}}\right] = [0,48; 0,52]$$

puisque $\frac{1}{\sqrt{n}} = \frac{1}{50} = 0,02$. Or $f = 0,46 \notin [0,48; 58]$. Nous pouvons donc conclure, avec une marge de 5% que l'hypothèse : « le dé est équilibré » n'est pas vraie.

Nous allons poursuivre l'étude de ce genre de raisonnements dans la section suivante.

10.3 Test statistiques

De manière générale, nous faisons face à la situation suivante : au sein d'une population, nous supposons que la proportion d'un certain caractère vaut $p \in [0;1]$. Nous sommes souvent amené à faire des suppositions sur cette proportion. Par exemple,

$$H_0 \,:\, p = \frac{1}{2} \quad \text{ou} \quad H_1 \,:\, p \neq \frac{1}{2}. \tag{10.3.1}$$

A partir d'un échantillon de taille n dans lequel le caractère en question est présent avec une fréquence valant f. Qu'est-il possible de décider ? Au vu des observations, devons-nous accepter l'hypothèse H_0 portant sur p ou la rejeter ? Dans tous les cas, quelle est l'erreur commise ?

Remarque. Implicitement, nous utilisons une loi binomiale dont l'épreuve de Bernoulli associée serait

$$S \,:\, \text{le caractère étudié apparait} \quad ; \quad \mathbb{P}(S) = p$$

et le nombre de répétitions indépendantes correspond à la taille de l'échantillon .

Voyons quelques exemples.

Exemple 10.3.1.

1. L'étude de l'équilibre d'une pièce

$$\textit{i.e.} \quad H_0 \,:\, p = \frac{1}{2} \quad \text{contre} \quad H_1 \,:\, p \neq \frac{1}{2}$$

   rentre dans le cadre décrit plus haut.

2. La victoire d'un candidat à une élection

$$\textit{i.e.} \quad H_0 \,:\, p > \frac{1}{2} \quad \text{contre} \quad H_1 \,:\, p < \frac{1}{2}$$

   rentre également dans ce cadre d'étude.

Afin d'obtenir une règle de décision, permettant de trancher lors d'un test, nous aurons besoin de la définition d'un « intervalle de fluctuation ».

Définition 10.3.1. *Dans le cadre décrit précédemment, un intervalle de fluctuation (avec un seuil d'erreur de 5%) d'une fréquence, sur un échantillon aléatoire de taille n est donné :*

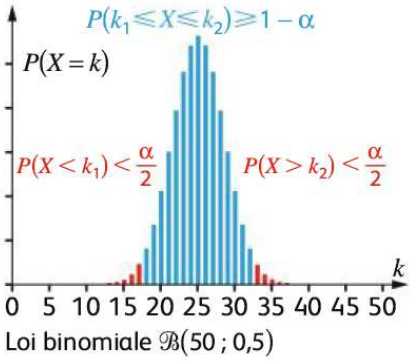
$$T_n = \left[\frac{k_1}{n}; \frac{k_2}{n}\right] \quad \textit{avec} \quad \begin{cases} k_1 \,:\, \textit{le plus petit entier tel que} & \mathbb{P}(X < k_1) < 0,025 \\ k_2 \,:\, \textit{le plus petit entier tel que} & \mathbb{P}(X > k_2) < 0,025 \end{cases}$$

où X désigne le nombre de fois où le caractère étudié apparait dans l'échantillon.

Remarque. Il est possible de remplacer la marge d'erreur de 5% par $\alpha \in]0;1[$. Dans ce cas, la définition de k_1 et k_2 est changée en

$$T_n = \left[\frac{k_1}{n}; \frac{k_2}{n}\right] \quad \text{avec} \quad \begin{cases} k_1 \,:\, \text{le plus petit entier tel que} & \mathbb{P}(X < k_1) < \frac{\alpha}{2} \\ k_2 \,:\, \text{le plus petit entier tel que} & \mathbb{P}(X > k_2) < \frac{\alpha}{2} \end{cases}$$

Graphiquement, les entiers k_1 et k_2 sont choisis de sorte que les aires en rouge soient inférieures à $\frac{\alpha}{2}$:



Nous pouvons alors énoncer la règle de décision suivante : laquelle qui apporte une solution au problème (10.3.1).

Proposition 36 (Règle de décision)**.** *Nous avons l'alternative suivante :*

1. *si $f \in T_n$ alors, avec une marge d'erreur de 5%, l'hypothèse H_0 n'est pas rejetée .*
2. *si $f \notin T_n$ alors, avec une marge d'erreur de 5%, l'hypothèse H_0 est rejetée .*

Remarque. D'un point de vue pratique, les **tests statistiques sont plutôt là pour rejeter des hypothèses plutôt que les accepter.** Cela revient à dire : si je modélise mon aléa pour une certaine loi de probabilité, est-ce que les **fréquences observées sont réalistes ou non ?** Comme mentionné plutôt, la marge d'erreur de 5% peut-être remplacée par $\alpha \in]0;1[$.

Voyons sur un exemple.

Exemple 10.3.2. Reprenons l'exemple de la pièce et testons

$$H_0 \,:\, p = \frac{1}{2} \quad \text{contre} \quad H_1 \,:\, p \neq \frac{1}{2}.$$

Supposons qu'après 50 lancers, nous ayons observé 19 piles. D'où $f = \frac{19}{50}$. A l'aide de la calculatrice nous déterminons la valeur de k_1 et de k_2 (associée à une loi binomiale de paramètre $n = 50$ et $p = \frac{1}{2}$). Nous trouvons

$$k_1 = 18 \quad \text{et} \quad k_2 = 32.$$

Par suite, $T_n = [\frac{18}{50}; \frac{32}{50}]$. Puisque $f \in T_n$, nous ne rejetons pas l'hypothèse H_0. Autrement dit, d'après les observations, il n'est pas aberrant de dire que la pièce est équilibrée (mais nous ne savons toujours pas si c'est le cas).

Exercices à traiter : 55, 56, 58 page 214 et 76 page 218 ; 75 et 78 page 218 en DM.