Chapitre 12

Échantillonnage et estimation

Objectifs du chapitre :

item	*références*	*auto évaluation*				
centrer et réduire une loi binomiale						
déterminer et utiliser un intervalle de fluctuation asymptotique						
déterminer et utiliser un intervalle de confiance						

1) Échantillonnage

1 - 1) Intervalle de fluctuation avec la loi binomiale

Il a été vu en classe de 1ère la notion de fluctuation d'échantillonnage.

Lorsqu'on modélise une situation, les résultats ne seront pas exactement égaux à ce qui avait été prévu ; c'est ce qu'on appelle la fluctuation d'échantillonnage.

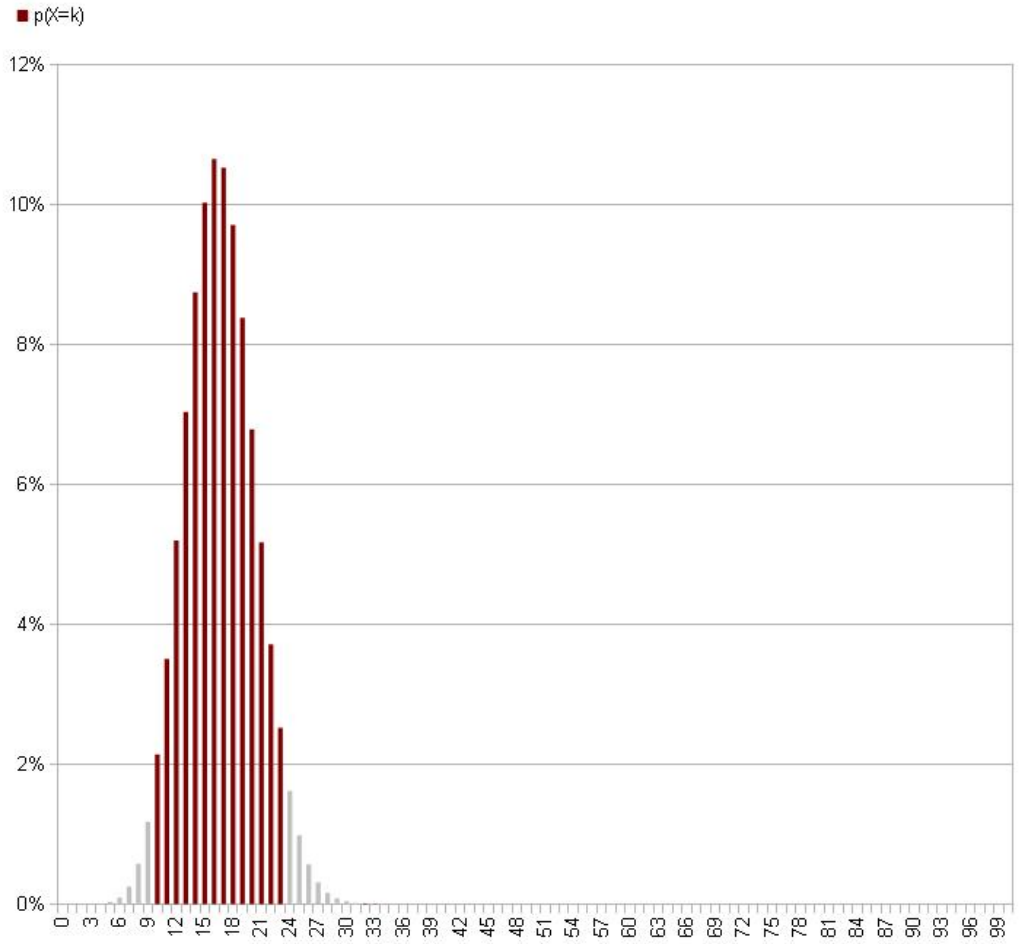
Par exemple, on peut modéliser le résultat du lancer d'un dé à six faces en disant que chaque face possède une chance sur six de sortir, autrement dit que chaque face a une probabilité de sortir égale à $\frac{1}{6}$.

Pour autant, sur six lancers, chaque face ne sort pas exactement une fois ...

Étudier la fluctuation d'échantillonnage, c'est déterminer un intervalle dans lequel les résultats peuvent être attendus, **en conformité avec le modèle choisi**.

En classe de 1ère, la loi binomiale a été utilisée pour mettre en place des intervalles de fluctuation au seuil de 95%, c'est-à-dire un intervalle regroupant les valeurs qui devraient théoriquement sortir dans 95% des cas.

Le graphique ci-dessous représente la loi binomiale de paramètres $n = 100$ et $p = \frac{1}{6}$, autrement dit la probabilité d'obtenir le n°1 par exemple lorsqu'on jette un dé à six faces.



L'intervalle $[10; 24]$ regroupe 95% des issues possibles : c'est ce qu'on appelle l'intervalle de fluctuation au seuil de 95%.

On peut raisonner en terme de **fréquence** : dans 95% des cas, la fréquence d'apparition du n°1 est comprise entre $\frac{10}{100}$ et $\frac{24}{100}$; on peut noter : $f \in [10\% \ ; \ 24\%]$.

Établir un intervalle de fluctuation à l'aide de la loi binomiale (en terme d'occurrence ou de fréquence) est long, nécessite de nombreux calculs. On va voir comment être plus efficace, en utilisant la loi normale centrée réduite.

1 - 2) Variable centrée réduite

Pour se rapprocher d'une loi normale, il est nécessaire que la variable aléatoire considérée soit :
– « **centrée** », ce qui veut dire que sont espérance mathématique est nulle ;
– « **réduite** », ce qui veut dire que sa variance (et donc son écart-type) est égal à 1.

On peut rappeler que :
* une loi binomiale $\mathcal{B}(n;p)$ a pour espérance mathématique np ;

* une loi binomiale $\mathcal{B}(n;p)$ a pour variance $np(1-p)$ et donc pour écart-type $\sqrt{np(1-p)}$;

* l'espérance est linéaire, c'est-à-dire que $E(aX+b) = aE(X) + b$;

* la variance possède la propriété suivante : $V(aX+b) = a^2V(X)$; en particulier, $V(b) = 0$.

Dans la suite, on considère une variable aléatoire X_n qui suit une loi binomiale $\mathcal{B}(n;p)$.

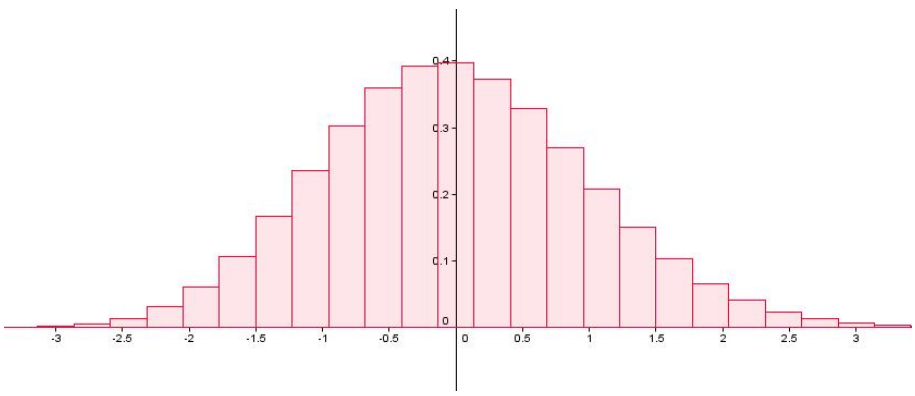
On « centre » la variable : $X_n - np$ est une variable aléatoire d'espérance nulle, car : $E(X_n - np) = E(X_n) - E(np) = np - np = 0$.

On « réduit » la variable : $\dfrac{X_n - np}{\sqrt{np(1-p)}}$ a pour variance 1, car :

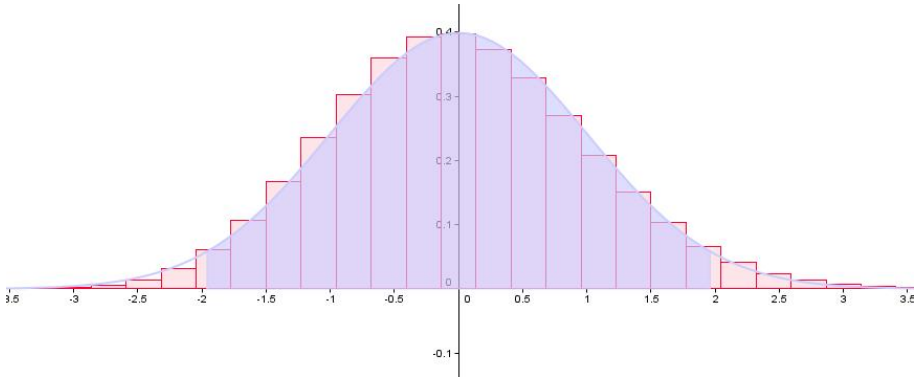
$$V\left(\frac{X_n - np}{\sqrt{np(1-p)}}\right) = \frac{1}{\sqrt{np(1-p)}^2}(V(X_n) - V(np)) = \frac{1}{\sqrt{np(1-p)}^2}(np(1-p) - 0) = 1$$

Ainsi, on a construit une variable aléatoire centrée réduite Z_n à partir de la variable X_n, avec $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$

En voici une représentation graphique (pour $n = 100$ et $p = \dfrac{1}{6}$) :



Cette représentation graphique montre « qu'on a envie » d'approximer cette variable la loi normale (centrée réduite) dont on sait déterminer les bornes regroupant 95% des valeurs :



La suite de ce cours va consister à formaliser cette idée : approcher une variable aléatoire centrée réduite issue d'une loi binomiale par la loi normale.

1 - 3) Propriété de la variable aléatoire fréquence F_n

Soit X_n une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n;p)$; on définit la variable aléatoire F_n par $F_n = \dfrac{X_n}{n}$; elle représente **la fréquence de succès** pour un schéma de Bernoulli de paramètre n et p.

L'intérêt de cette variable est qu'elle est comprise en 0 et 1, alors qu'une variable du type X_n est comprise entre 0 et n.

La variable aléatoire F_n ne suit pas une loi binomiale ; néanmoins, sa loi de probabilité et la représentation graphique de cette loi se déduisent directement de celles de X_n.

Théorème :

Si X_n est une variable aléatoire suivant une loi binomiale $\mathcal{B}(n;p)$, alors, pour tout α dans $]0;1[$, on a :

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha$$

où I_n désigne l'intervalle $\left[p - u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

Commentaires :

Il faut bien se remettre en tête les notations vues dans le chapitre 14, en particulier la notation u_α.

Lorsqu'on choisit une valeur α, le nombre noté u_α est tel que : $P(-u_\alpha \leqslant Y \leqslant u_\alpha) = 1-\alpha$, où Y est une variable aléatoire suivant la loi normale centrée réduite.

On peut à ce propos rappeler que : $P(-u_\alpha \leqslant Y \leqslant u_\alpha) = \displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$

Démonstration - BAC :

A partir d'une variable aléatoire X_n suivant $\mathcal{B}(n;p)$, on pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$

Z_n est une variable aléatoire centrée et réduite.

Nous pouvons débuter la démonstration proprement dite, en transformant les écritures pour faire apparaître cette variable aléatoire centrée réduite :

$$\begin{aligned}
\frac{X_n}{n} \in I_n \quad &\Leftrightarrow \quad p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\
&\Leftrightarrow \quad np - u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}} \leqslant X_n \leqslant np + u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}} \\
&\Leftrightarrow \quad np - u_\alpha \sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha \sqrt{np(1-p)} \\
&\Leftrightarrow \quad -u_\alpha \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant u_\alpha
\end{aligned}$$

D'après le théorème de Moivre Laplace,

$$\lim_{n\to+\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = \int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

Or, par définition des valeurs $-u_\alpha$ et u_α, $\displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt = 1 - \alpha$

Ainsi, $\displaystyle\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1 - \alpha$

1 - 4) Intervalle de fluctuation asymptotique

Le paragraphe précédent permet de déterminer efficacement un intervalle regroupant environ 95% des valeurs pour la variable aléatoire Z_n, en approximant cette variable aléatoire par la loi normale.

Défintion :

L'intervalle $\left[p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est un **intervalle de fluctuation asymptotique au seuil de confiance** $1-\alpha$ de la variable aléatoire F_n qui, à tout échantillon de taille n, associe la fréquence obtenue.

Commentaires :

* Cet intervalle contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand.

* Cette approximation est valable dès que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Exemple :

En reprenant l'exemple du jet de dé, en cherchant un intervalle de fluctuation au seuil de 95 %, on sait que 95 % des valeurs se trouvent pour une loi normale centrée réduite dans l'intervalle $[-1,96\ ;\ 1,96]$; ainsi, environ 95 % des valeurs de la variable Z_n se trouvent dans cet intervalle.

Or, $Z_n \in [-1,96; 1,96] \Leftrightarrow \dfrac{X_n}{n} \in \left[p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\ ;\ p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

Avec les valeurs $n = 100$ et $p = \dfrac{1}{6}$, on trouve un intervalle égal à : $\boxed{[0,094\ ;\ 0,24]}$, soit encore $[9,4\%\ ;\ 24\%]$

On remarque que cet intervalle est très proche de celui trouvé par la méthode de 1ère, et il n'a nécessité que peu de calculs.

Cas particulier important :

Comme dans l'exemple précédent, on utilise très souvent l'**Intervalle de fluctuation asymptotique au seuil de confiance de 95 %** :

$$I_{\text{fluctuation à 95 \%}} = \left[p - 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\ ;\ p + 1,96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

Conditions d'application : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

2) Prise de décision à partir d'un échantillon

On met en place un modèle concernant une population, en particulier **une hypothèse sur une proportion dans cette population**.

Propriété :

On considère une population dans laquelle on **suppose** que la proportion d'un caractère est p.

On **observe** f comme fréquence de ce caractère dans un échantillon de taille n.

On fait l'hypothèse : « la proportion de ce caractère dans la population est p. »

En notant I l'intervalle de fluctuation de la fréquence à 95% dans les échantillons de taille n, alors, la **règle de décision** est la suivante :

- si $f \notin I$: on **rejette** l'hypothèse, le risque de se tromper étant de 5% ;
- si $f \in I$: rien ne permet de rejeter l'hypothèse, donc on l'**accepte** (au risque de se tromper ... sans pouvoir estimer l'erreur)

3) Estimation d'une proportion

Le problème de l'estimation peut être considéré comme le problème « inverse » de celui de l'échantillonnage : à partir de la fréquence f observée sur un échantillon, on cherche à estimer la proportion p correspondante dans la population toute entière. C'est le problème que l'on cherche à résoudre en réalisant des sondages.

On rappelle la formule vue en seconde donnant une approximation de l'intervalle **de fluctuation** au seuil de 95 % de la fréquence d'un caractère pour un échantillon de taille n, la proportion de ce caractère (ou la probabilité de trouver ce caractère) au sein de la population totale étant égale à p : $\left[p - \dfrac{1}{\sqrt{n}} \ ; \ p + \dfrac{1}{\sqrt{n}}\right]$.

A partir de l'inégalité $p - \dfrac{1}{\sqrt{n}} \leqslant f$, on obtient : $p \leqslant f + \dfrac{1}{\sqrt{n}}$

Et à partir de l'inégalité $f \leqslant p + \dfrac{1}{\sqrt{n}}$, on obtient : $f - \dfrac{1}{\sqrt{n}} \leqslant p$

Ainsi, un travail sur l'intervalle de fluctuation vu en seconde permet de justifier la définition suivante :

Définition :

Soit f la fréquence observée d'un caractère dans un échantillon de taille n extrait d'une population dans lequel la proportion de ce caractère est p.

Alors l'intervalle $\left[f - \dfrac{1}{\sqrt{n}} \ ; \ f + \dfrac{1}{\sqrt{n}}\right]$ est un **intervalle de confiance de la proportion p au niveau de confiance 95%**.

Commentaires :

* **On utilise cet intervalle dès que $n \geqslant 30$, $nf \geqslant 5$ et $n(1-f) \geqslant 5$.**

* Cet intervalle est la réalisation, à partir d'un échantillon, d'un intervalle aléatoire contenant la proportion p avec une probabilité supérieure ou égale à 0,95.

* Dans certains domaines, on utilise un intervalle de confiance plus précis, au niveau de confiance 95% : $\left[f - 1,96\dfrac{\sqrt{f(1-f)}}{\sqrt{n}} \ ; \ f + 1,96\dfrac{\sqrt{f(1-f)}}{\sqrt{n}}\right]$

Exemple :

Un sondage est réalisé auprès de 2500 personnes. 49 % des sondés déclarent qu'ils voteront pour Monsieur X. On peut estimer la proportion de personnes qui voteront réellement pour Monsieur X (les conditions d'utilisation de la formule précédente sont vérifiées) :

$$\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right] = \left[0,49 - \frac{1}{\sqrt{2500}} \ ; \ 0,49 + \frac{1}{\sqrt{2500}}\right] = [0,47 \ ; \ 0,51]$$

On n'est pas complètement sûr que Monsieur X va perdre les élections ...