CHAPITRE XV

Échantillonnage

Dans une population, on souhaite étudier un certain caractère (par exemple, dans une population d'humains, on peut étudier les personnes porteuses d'un certain gène).
Si la population est trop nombreuse, on ne peut pas étudier tous les individus donc on prélève un *échantillon* de taille n pour étudier ce caractère.
Il y a deux situations possibles :

1. Si on connaît la proportion exacte p de ce caractère dans la population alors on étudiera les fluctuations de l'échantillon par rapport à la population totale (voir I de ce chapitre).
2. Si on ne connaît pas la proportion exacte p de ce caractère dans la population, on essaiera de faire une estimation de p à partir de l'échantillon (voir II de ce chapitre).

Dans tout ce chapitre, on supposera que $n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$.

I/ Fluctuations d'échantillonnages

1. Variable aléatoire fréquence

Définition I.1

Dans une population, on étudie un caractère qui a une proportion p. On effectue l'expérience aléatoire qui consiste à tirer un échantillon de taille n de cette population et on note X la v.a. qui compte le nombre d'individus qui possèdent ce caractère.
On sait alors que X suit la loi binomiale de paramètres n et p.
La fréquence de ce caractère dans l'échantillon est la variable aléatoire F définie par :

$$F = \frac{X}{n}$$

2. Intervalles de fluctuations asymptotiques

Si on tire plusieurs échantillons dans la population, la fréquence ne sera pas toujours la même à chaque fois : on dira que la fréquence fluctue. Un intervalle de fluctuations I est un intervalle auquel la variable aléatoire F appartient avec une probabilité de 95%.

Définition I.2

Un *intervalle de fluctuations asymptotique* au seuil de 95% de la variable aléatoire F est un intervalle I tel que $P(F \in I) \simeq 95\%$.

Théorème I.3

Un intervalle de fluctuations asymptotique de F est l'intervalle :

$$I = \left[p - 1,96\sqrt{\frac{p(1-p)}{n}} ; p + 1,96\sqrt{\frac{p(1-p)}{n}} \right]$$

(où p est la proportion du caractère dans la population et n est la taille de l'échantillon).

□ ***Exemple I.1.*** En Écosse, on sait que 13% de la population est rousse. On choisit au hasard un échantillon de 50 Écossais et on note F la fréquence de personnes rousses dans cet échantillon.
Calculer un intervalle de fluctuations asymptotique de F au seuil de 95% et interpréter. → À rédiger

3. Test d'hypothèse et prise de décision

Parfois, on peut être amené à faire une hypothèse sur la valeur de la proportion p d'un caractère dans une population. Pour tester cette hypothèse, on suit le protocole suivant :

- On tire un échantillon de taille n dans la population.
- On calcule la fréquence observée f de ce caractère dans l'échantillon.
- On calcule un intervalle de fluctuations asymptotiques I à l'aide de p et n.
- Puis on utilise la **règle de décision** suivante :
  - Si f appartient à l'intervalle de fluctuations I alors on accepte l'hypothèse sur la valeur de p ;
  - Sinon, on rejette cette hypothèse avec un risque de 5% de chances de se tromper.

☐ ***Exemple I.2.*** Une entreprise qui fabrique des galettes décide de mettre des fèves en or dans une partie de ses galettes. Cette entreprise affirme que 15% de ses galettes ont une fève en or.
Pour vérifier cette affirmation, la direction de la répression des fraudes commande 50 galettes et constate que 2 galettes seulement possèdent une fève en or. Cette entreprise doit-elle être sanctionnée pour publicité mensongère ? Justifier votre réponse. → À rédiger

II/ Intervalles de confiance

Dans ce paragraphe, on se place dans la situation où la proportion p du caractère n'est pas connue et on cherche à l'estimer.

Définition II.1

Un *intervalle de confiance* de la proportion p au niveau de confiance de 0,95 est la réalisation à partir d'un échantillon d'un intervalle aléatoire contenant p avec une probabilité supérieure ou égale à 95%.

Ce qu'il faut retenir de cette définition obscure est que se donner un intervalle de confiance revient à se donner une méthode qui donne un intervalle qui contient la proportion p dans 95% des cas.

Théorème II.2

Si f est la fréquence observée du caractère sur un échantillon de taille n alors un intervalle de confiance au niveau de confiance de 0,95 est :

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

Interprétation : Cela signifie que si on tirait 100 échantillons de taille n et qu'à chaque fois on calculait l'intervalle de confiance $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$, alors la proportion réelle p du caractère appartiendrait à environ 95 de ces intervalles.

☐ ***Exemple II.1.*** Une semaine avant le 2nd tour d'une élection municipale, un sondage est effectué sur 1024 personnes choisies au hasard parmi les 42 821 inscrits. De ce sondage ressort que 532 des personnes interrogées ont déclaré vouloir voter pour le maire sortant.
Le maire sortant peut-il envisager d'être réélu ? Justifier votre réponse. → À rédiger

Propriété II.3

L'amplitude d'un intervalle de confiance est $\frac{2}{\sqrt{n}}$.

☐ ***Exemple II.2.*** On veut estimer la proportion p de personnes immunisées contre un certain virus parmi la population d'une ville. On prélèvre un échantillon aléatoire de 500 personnes parmi cette population. La population est suffisamment importante pour assimiler ce prélèvement à un tirage au hasard avec remise.

1. Après analyse, on dénombre 241 personnes immunisées contre ce virus parmi les 500 personnes de cet échantillon. Donner un intervalle de confiance de la proportion de personnes immunisées contre ce virus parmi la population de la ville, avec un niveau de confiance de 95%.
2. Quelle est la taille minimale de l'échantillon qui aurait permis d'obtenir un intervalle de confiance à 95% d'amplitude inférieure ou égale à 0,05 ? → À rédiger