Chapitre 13

Échantillonnage

13.1 Intervalle de fluctuation et prise de décision

Propriété 13.1. *Soit un caractère d'une population dont la proportion est p. Lorsque la taille n des échantillons est telle que $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, alors il y a au moins 95% des échantillons au sein desquels la fréquence f du caractère appartient à l'intervalle*

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

C'est ***l'intervalle de fluctuation*** *à 95%.*

Démonstration. Admis. □

Remarques :

- La constante 1,96 découle d'une loi de probabilité appelée loi normale et est liée à la précision de l'intervalle : 95%.
- Les intervalles de fluctuations permettent la prise de décision, notamment sur la validité ou non d'hypothèses. En effet, on peut effectuer des hypothèses sur la proportion de certains caractères au sein d'une population ou la représentativité d'un échantillon vis à vis de cette population.

Exemples : Il y a un peu près 51% de femmes actuellement en France. On souhaite savoir si l'assemblée nationale respecte la parité homme/femme. Pour cela, on peut regarder si la fréquence de femmes au sein de l'assemblée nationale est dans l'intervalle de fluctuation à 95%. En effet, on peut considérer que l'assemblée nationale constitue un échantillon de la population française. Si la fréquence de femmes au sein de l'assemblée nationale n'est pas dans l'intervalle de fluctuation, on pourra considérer que celle-ci n'est pas représentative de la population française.

L'assemblée nationale compte actuellement 577 membres, on a donc un échantillon de taille $n = 577$. L'intervalle de fluctuation à 95% associé est donc

$$\left[0,51 - 1,96\sqrt{\frac{0,51 * 0,49}{577}}\,;0,51 + 1,96\sqrt{\frac{0,51 * 0,49}{577}}\right],$$

soit, en environ,

$$[0,47\,;0,55].$$

Il y a actuellement 224 femmes dans l'assemblée nationale, soit une fréquence d'environ $0,39 = 39\%$ des députés. Comme $0,39 \notin [0,47\,;0,55]$, on peut en déduire que l'assemblée nationale ne représente pas la population en matière de parité homme/femme.

13.2 Estimation d'une proportion

Propriété 13.2. *On note f la fréquence observée d'un caractère dans un échantillon de taille n issu d'une population au sein de laquelle la proportion de ce caractère est p. Alors au moins 95% des intervalles de la forme*

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

contiennent p. On les appelle ***intervalle de confiance.***

Démonstration. Admis. □

Remarque :

- p n'appartient pas forcément à cet intervalle. On peut juste dire si l'on avait un très grand nombre d'échantillon de taille n, alors p appartiendrait à cet intervalle dans 95% des cas.
- Les intervalles de confiances sont très utilisés pour effectuer les sondages.

Exemples : Au sein d'une population, on réalise une étude pour savoir la proportion p de ladite population inquiète vis à vis d'un effondrement de la civilisation dans les années à venir. Pour cela on interroge 1000 personnes à ce sujet. On dénombre 23 personnes inquiètes de l'effondrement, soit une fréquence de $f = 0,023$. On peut ainsi déterminer une estimation de la proportion p grâce à intervalle de confiance à 95% :

$$\left[0,023 - 1,96\sqrt{\frac{0,023(1-0,023)}{1000}}\,;0,023 + 1,96\sqrt{\frac{0,023(1-0,023)}{1000}}\right],$$

soit

$$[0,014\,;0,032].$$

On en déduit que $p \in [1,4\%\,;3,2\%]$ avec un seuil de confiance de 95%.

Attendus et savoir-faire :

- Déterminer un intervalle de fluctuation et prendre une décision à partir de celui-ci.
- Déterminer un intervalle de confiance et en déduire un encadrement d'une proportion.

13.3 Exercices

Exercice 13.1. On considère une population de taille n ayant un caractère en proportion p Dans chaque cas, déterminer si les hypothèses permettant de donner un intervalle de fluctuation sont vérifiées et le cas échéant donner l'intervalle de fluctuation.

1. $n = 1000$ et $p = 0,36$.
2. $n = 135$ et $p = 0,15$.
3. $n = 10$ et $p = 0,67$.
4. $n = 53$ et $p = 0,71$.

Exercice 13.2. [Python] Recopier et compléter la fonction ci-dessous afin qu'elle nous donne les bornes de l'intervalle de fluctuation en prenant comme arguments la taille de l'échantillon n et la proportion p.

```
def Fct_Bornes_Inter_Fluctu(...,...) :
    borne_inf=...
    borne_sup=...
    return borne_inf,borne_sup
```

Exercice 13.3. [Épidémiologie] Le réseau Sentinelles modélise un niveau de base de l'incidence de syndromes grippaux, i.e. le nombre de syndromes grippaux attendus une semaine donnée pour 100 000 habitants en l'absence d'épidémie. En effet, tout au long de l'année et même lorsque les virus de la grippe ne circulent pas, des syndromes grippaux sont observés par des médecins. L'incidence attendue hors épidémie est estimée chaque semaine par un modèle et assortie d'un intervalle de fluctuation. Cet intervalle indique entre quelles valeurs doit se trouver l'incidence observée avec une certaine probabilité s'il n'y a pas d'épidémie ; la borne supérieure de cet intervalle définit le seuil épidémique à partir duquel on peut considérer qu'il y a épidémie.

1. Pour une certaine semaine donnée, le niveau de base d'incidence de la grippe dans le pays était de 192 cas pour 100 000 habitants. Déterminer l'intervalle de fluctuation associé à ce niveau de base d'incidence.
2. En déduire le seuil épidémique pour cette semaine là.
3. Une région 4 985 000 habitants compte 11964 cas de grippe cette semaine là, le seuil épidémique est-il passé ?

Exercice 13.4. [Python] Recopier et compléter la fonction ci-dessous afin qu'elle nous donne les bornes de l'intervalle de confiance en prenant comme arguments la taille de l'échantillon n et la fréquence f.

```
def Fct_Bornes_Inter_Conf(...,...) :
    borne_inf=...
    borne_sup=...
    return borne_inf,borne_sup
```

Exercice 13.5. [Sondages] En période d'élections les sondages sont nombreux et servent à construire des récits autour de leurs protagonistes. Dans le Pays Fictif (c'est son nom), trois partis politiques se distinguent en vue d'accéder au pouvoir et de gouverner dans leurs intérêts : le C'était Mieux Avant, le Ne Changeons Rien et l'Illusion du Progrès. Jusqu'ici, c'est le Ne Changeons Rien qui était en tête des intentions de votes avec son candidat F. Filou :

1. Ne Changeons Rien : 22% ;
2. L'Illusion du Progrès : 21% ;
3. C'était Mieux Avant : 19% ;
4. le reste n'étant qu'un ramassis de petits candidats et de dangereux abstentionnistes faisant le jeu du C'était Mieux Avant.

Toutefois, le candidat de l'Illusion du Progrès, E. Monarc, a fait un beau discours hier et ce matin un nouveau sondage effectué auprès de 1000 personnes donne :

1. L'Illusion du Progrès : 23% ;
2. Ne Changeons Rien : 22% ;
3. C'était Mieux Avant : 18% ;
4. le reste n'étant qu'un ramassis de petits candidats et de dangereux abstentionnistes faisant le jeu du C'était Mieux Avant.

Les médias font alors les louanges d'E. Monarc et de son discours, de comment il a pu retourner la situation grâce à son génie. Cependant, il ne s'agit que d'un sondage et il faut se demander si ses résultats sont transposables à l'ensemble de la population. On note p_P la cote de popularité de l'Illusion du Progrès et p_R celle du Ne Changeons Rien.

1. Donner un intervalle de confiance de p_P pour les anciens sondages.
2. Donner un intervalle de confiance de p_P pour le nouveau sondage. Qu'en déduisez-vous ?
3. Donner un intervalle de confiance de p_R pour le nouveau sondage. Le comparer avec celui de p_P. Qu'en déduisez-vous ?

Exercice 13.6. [Zombies] Afin de se prémunir d'une attaque de castors-zombies, on souhaite s'intéresser à l'efficacité des pistolets laser comme moyen de les combattre. Pour cela, on veut réaliser une expérience sur un échantillon représentatif de castors-zombies afin d'obtenir une idée du taux de succès du pistolet laser en combat. Le hic : il existe des castors-zombies-jedi, lesquels peuvent se défendre contre les pistolets laser ; il est donc important que l'échantillon soit représentatif de leur présence au sein des castors-zombies. La proportion de castors-zombies-jedi au sein de la population totale est de 1%. On sélectionne un échantillon de 1000 castors-zombies.

1. Quel est l'intervalle de fluctuation à 95% associé à cet échantillon ? On précisera les quantités utilisées.
2. Parmi l'échantillon des mille castors, on dénombre 17 castors-zombies-jedi, peut-on considérer l'échantillon comme représentatif ? Justifier.
3. On a testé les pistolets laser sur l'échantillon des mille castors-zombies, ceux-ci se sont avérés efficaces dans 64% des cas. Quel est l'intervalle de confiance à 95% associé à cet échantillon ? On précisera les quantités utilisées.
4. Est-il possible que la proportion de castors-zombies vulnérables au pistolet laser soit $p = 0,4$? Justifier.
5. Quel nom donneriez-vous à un film avec des castors-zombies-jedi ?

Exercice 13.7. [Python] On considère le programme en Python ci-dessous. *Indications :* random est la librairie de Python contenant les outils pour l'aléatoire et random.randint(a,b) est une fonction prenant au hasard un nombre entier entre a et b.

```
import random

succes=0
for simu in range(100):
   if random.randint(0,1)==0:
       succes=succes+1
print(succes/100)
```

1. Quel est le rôle de la variable succes ?
2. Que fait ce programme ?
3. Ce programme convient-il pour simuler :
   (a) 100 lancers d'une pièce équilibrée dont on compterait le nombre de pile ?
   (b) 100 tirages avec remise d'une boule dans une urne contenant 5 boules rouges et 15 bleues dont on compterait les boules rouges ?

Exercice 13.8. [Python]

1. En France, la proportion de personne ayant les cheveux blonds est de 10%. Que fait le programme ci-dessous ?

   ```
   import random

   nb_individus=50
   succes=0
   for simu in range(nb_individus):
      if random.randint(1,10)==1:
          succes=succes+1
   print(succes)
   ```

2. Modifier ce programme pour qu'il simule la constitution d'un échantillon de 100 individus et affiche le nombre d'entre eux ayant les cheveux brun foncé sachant que la proportion d'individus en France ayant les cheveux brun foncé est de $2,5\%$.

Exercice 13.9. [Tableur, Médecine,*]** En 2009, des chercheurs annoncent un résultat « statistiquement significatif » dans la recherche d'un vaccin contre le sida. Des personnes séronégatives ont reçu (sans pourvoir faire la différence) soit un placebo, soit le vaccin expérimental. Quatre ans après, sur les 8198 personnes ayant reçu le placebo, 74 ont été infectées par le virus ; sur les 8197 personnes vaccinées, 51 ont été infectés.

Le groupe placebo permet d'estimer la proportion habituelle d'infections à $p = \dfrac{74}{9198} \simeq 0,009$.

1. Déterminer l'intervalle de fluctuation à 95% du nombre d'infectés. Comparer avec les résultats du vaccin.
2. On va maintenant effectuer des simulations à l'aide du tableur LibreOffice.
   (a) Dans la case A1 du tableur, que permet de simuler l'instruction : =SI(ALEA()<0,009 ;"Infecté" ;"Non infecté") ?
   (b) On recopie cette formule 8197 fois cette instruction jusqu'à la case A8197 et dans la A8198, on entre l'instruction =NB.SI(A1 :A8197 ;"Infecté"). À quoi correspond le nombre affiché ?
   (c) En étirant les instructions précédentes jusqu'à la colonne 100, simuler 100 échantillon de 8197 personnes et compter les nombres d'infectés pour chacun d'entre eux.
   (d) En utilisant l'instruction =SI(A8197<=51 ;1 ;0), compter le nombre d'échantillons ayant moins de 51 infectés.
   (e) Que signifie l'expression « statistiquement significatif » pour le vaccin expérimental ?