Chapitre 18

Échantillonnage

18.1 Échantillon

Définition 18.1. *Soit un caractère d'une population présent dans cette dernière avec une proportion p.*

- *On appelle* ***échantillon*** *de taille n de cette population toute liste de n individus pris au hasard, indépendamment les uns des autres et de façon identiquement distribuée : chaque individu possède le caractère avec une probabilité p.*
- *On appelle* ***fréquence*** *observée f du caractère est la fréquence d'apparition du caractère dans l'échantillon.*

Remarques :

- Du point de vue de la définition, fréquence et proportion se calculent de la même façon. Toutefois, la proportion fait plutôt référence à la population dans son ensemble tandis que la fréquence fait référence à l'échantillon. D'une certaine façon, la proportion est théorique et la fréquence est empirique ou expérimentale.
- Il est possible de simuler informatiquement des échantillons et de calculer la fréquence d'apparition d'un caractère dans ceux-ci ou alors d'essayer d'estimer la proportion du caractère au sein de la population générale.

Exemple :

- En France, il y a 51% de femmes. Un échantillon de 100 personnes pour ce caractère est donc une liste de 100 personnes prises au hasard dans la population, de façon indépendante, et avec une probabilité de 0,51 pour chaque personne d'être une femme.
- Si dans un échantillon de 100 personnes, on a 55 femmes, la fréquence observée sera donc $f = \dfrac{55}{100} = 0,55$; on voit qu'elle diffère de la proportion de la population globale.

— Un programme Python permettant de calculer la fréquence de femmes au sein d'un échantillon de 100 personnes est celui ci-dessous. `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

nb_individus = 100
nb_femmes = 0

for simu in range(nb_individus) :
   if random.randint(1,100) <= 51 :
       nb_femmes = nb_femmes + 1

freq_femmes = nb_femmes / nb_individus

print(freq_femmes)
```

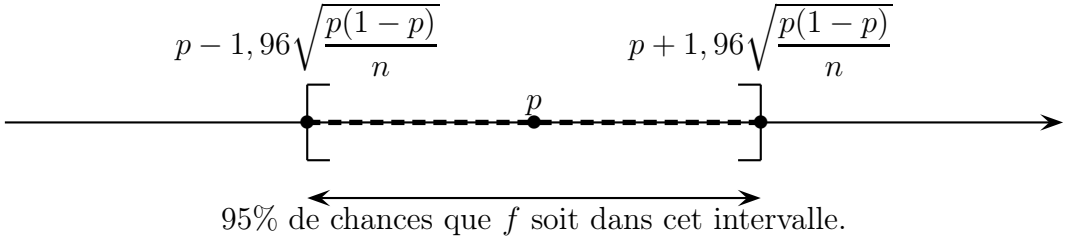
Exercices : 18.1 et 18.2.

18.2 Intervalle de fluctuation et prise de décision

Proposition 18.1. ***[Admise]*** *Soit un caractère d'une population dont la proportion est p. Lorsque la taille n des échantillons est telle que $n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$, 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%.



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

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

soit, en environ,

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

Avec une assemblée nationale comportant 224 femmes, on a 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.

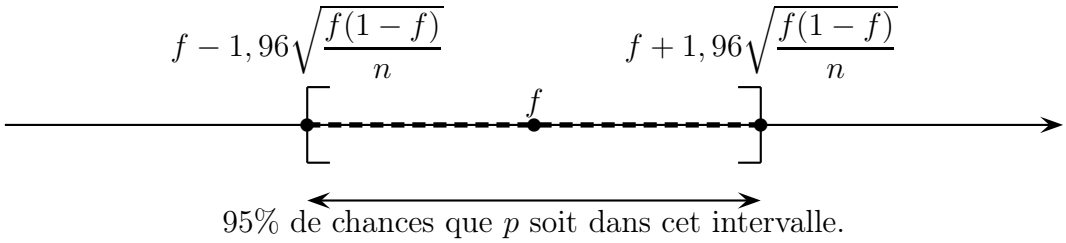
Exercices : 18.3 à 18.6.

18.3 Estimation d'une proportion

Proposition 18.2. ***[Admise]*** *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 ***intervalles de confiance****.*



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 1 000 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% :

$$\left[0,023 - 1,96\sqrt{\frac{0,023(1-0,023)}{1\,000}}\,;0,023 + 1,96\sqrt{\frac{0,023(1-0,023)}{1\,000}}\right],$$

soit

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

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

Exercices : 18.7 et 18.8.

18.4 Capacités attendues

— Lire et comprendre une fonction Python renvoyant le nombre ou la fréquence de succès dans un échantillon de taille n pour une expérience aléatoire à deux issues.

— Observer la loi des grands nombres à l'aide d'une simulation sur Python ou tableur.

— Simuler N échantillons de taille n d'une expérience aléatoire à deux issues. Si p est la probabilité d'une issue et f sa fréquence observée dans un échantillon, calculer la proportion des cas où l'écart entre p et f est inférieur ou égal à $\frac{1}{\sqrt{n}}$.

18.5 Exercices

18.5.1 Progresser

Échantillon

Exercice 18.1. [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 18.2. [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 $2,5\%$.

Intervalle de fluctuation et prise de décision

Exercice 18.3. 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. $n = 1\,000$ 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 18.4. [Python] 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 bornes_inter_fluctu(...,...) :

    borne_inf = ..................................................

    borne_sup = ..................................................

    return borne_inf, borne_sup
```

Exercice 18.5. [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 11 964 cas de grippe cette semaine là, le seuil épidémique est-il passé ?

Exercice 18.6. [Tableur] Le but de cet exercice est de simuler à l'aide du tableur 100 échantillons de 50 lancers d'une pièce de monnaie truquée : cette pièce à une probabilité de 0,6 de tomber sur pile.

1. **Détermination de l'intervalle de fluctuation à 95%.**
   (a) Écrire n dans la case `A1` et dans la case `B1` sa valeur : 50.
   (b) Écrire p dans la case `A2` et dans la case `B2` sa valeur : 0,6.
   (c) Écrire borne_inf dans la case `A3` et dans la case `B3` la formule permettant de calculer la borne inférieure de l'intervalle de fluctuation.
   (d) Écrire borne_sup dans la case `A4` et dans la case `B4` la formule permettant de calculer la borne supérieure de l'intervalle de fluctuation.

2. **Simulation d'un échantillon.**
   (a) Dans la case `A10` du tableur, que permet de simuler l'instruction ci-dessous ?

   ```
   =SI(ALEA()<=$B$2;"Pile";"Face")
   ```

   (b) On recopie cette formule 50 fois cette instruction jusqu'à la case `A60` et dans la `A8`, on entre l'instruction `=NB.SI(A10:A60;"Pile")`. À quoi correspond le nombre affiché ?
   (c) Entrer dans la case `A7` une formule permettant de calculer le nombre de piles obtenus.
   (d) Dans la case `A5`, écrire la formule renvoyant 1 si la fréquence observée est dans l'intervalle de fluctuation et 0 sinon :

   ```
   =SI(ET($B$3<=A7;A7<=$B$4);1;0)
   ```

3. **Simulations de 100 échantillons.**
   (a) En étirant les lignes `A5` à `A60` jusqu'à la colonne `CV`, simuler 100 échantillon de 50 lancers.
   (b) À l'aide de la fonction `SOMME`, déterminer dans la case `D1` le nombre d'échantillons dont la fréquence est dans l'intervalle de fluctuation. En a-t-on au moins 95% ?
   (c) Changer les formules entrées pour les bornes inférieures et supérieures pour $p - \dfrac{1}{\sqrt{n}}$ et $p + \dfrac{1}{\sqrt{n}}$. Combien d'échantillons sont cette fois-ci dans l'intervalle de confiance ?

Estimation d'une proportion

Exercice 18.7. [Python]

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 bornes_inter_conf(...,...) :

    borne_inf = ..................................................

    borne_sup = ..................................................

    return borne_inf, borne_sup
```

Exercice 18.8. [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 Changeons Tout et l'Illusion du Progrès. Jusqu'ici, c'est l'Illusion du Progrès qui est en tête des intentions de vote avec son candidat E. Monarc. Il devance les candidats du C'était Mieux Avant et du Changeons Tout, respectivement M. La Peine et J.L. Vieuronchon.

1. L'Illusion du Progrès : 20% ;
2. C'était Mieux Avant : 19% ;
3. Changeons Tout : 18% ;
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 1 000 personnes donne :

1. L'Illusion du Progrès : 21% ;
2. C'était Mieux Avant : 18% ;
3. Changeons Tout : 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 prendre l'ascendant sur ses adversaires grâce à son génie. Cependant, il ne s'agit que d'un sondage et il faut se demander si ces résultats sont transposables à l'ensemble de la population. On note p_P la cote de popularité de l'Illusion du Progrès, p_A celle du C'était Mieux Avant et p_T celle du Changeons Tout.

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_T et de p_A pour le nouveau sondage. Le comparer avec celui de p_P. Qu'en déduisez-vous ?