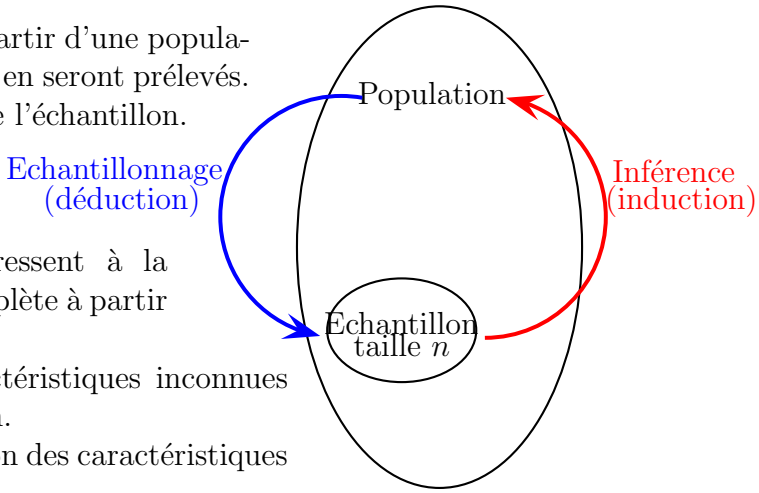
Echantillonnage - Statistique inférentielle

I - Introduction

L'échantillonnage est l'étude des liens existant entre les paramètres (moyenne ou fréquence) des échantillons issus d'une population et les paramètres de la population complète.
L'échantillonnage statistique consiste à prédire, à partir d'une population connue les caractéristiques des échantillons qui en seront prélevés.
On parle aussi de déduction des caractéristiques de l'échantillon.



Inversement, les statistiques inférentielles s'interessent à la détermination des paramètres de la population complète à partir de ceux d'un échantillon.
L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir de celles d'un échantillon.
On parle aussi d'induction, ou encore d'extrapolation des caractéristiques à l'ensemble de la population.

Exemples de problèmes et questions

1. **Pile ou face truqué ?**
   J'ai joué à pile ou face avec un ami 10 fois de suite. J'ai parié à chaque fois sur pile, et ai perdu 8 fois sur les 10.
   Puis-je raisonnablement penser que sa pièce était truquée ?

   Même question, si en poursuivant la partie sur 100 lancers, j'ai perdu 80 fois.

   Même question, si en poursuivant la partie sur 1000 lancers, j'ai perdu 800 fois.

2. **Influence d'une suine à proximité**
   Une usine chimique est venue s'implanter près d'une ville il y a 3 ans. Pendant ces 3 ans sont nés dans cette ville 132 enfants dont 52 garçons.

   Peut-on considérer que l'usine a eu un impact sur les naissances ?

3. **Parité homme-femme**
   Deux entreprises A et B recrutent sur un même bassin d'emploi où il y a autant d'hommes que de femmes. L'entreprise A emploie 60 personnes, dont 26 femmes, tandis que l'entreprise B emploie 1050 personnes, dont 480 femmes.

   Ces deux entreprises respectent-elles la parité homme-femme ?

4. a. **Echantillonnage...**
   Au cours d'une consulation électorale, le candidat A a recueilli 55% des voix.
   Quelle est la probabilité d'avoir, lors d'un sondage préélectoral réalisé sur un échantillon aléatoire de 100 personnes parmi les électeurs, moins de 50% d'intentions de vote pour le candidat A.

b. **...et sondage** On effectue un sondage préélectoral sur un échantillon "représentatif" de 100 personnes. Le pourcentage calculé sur cet échantillon des intentions de vote pour le candidat A est de 55%.

Peut-on en déduire que, si lors des élections tous les électeurs suivent leur intention de vote du jour du sondage, le candidat A sera élu ?

5. **Contrôle destructif : dimensionnement d'un sondage**

Une usine produit 10 000 pièces par jour. Pour vérifier la bonne fabrication de ces pièces, on utilise sur un certain nombre d'entre elles un contrôle destructif : les pièces sélectionnées sont détruites puis analysées, et on détermine ainsi avec certitude si celles-ci possédaient un défaut ou non.

a. De combien de pièces doit-être formé l'échantillon prélevé pour estimer le pourcentage de pièces défectueuses à la sortie de l'usine ?

b. On connaît le taux théorique de pièces défectueuses (fourni par les constructeurs des machines de l'usine). Peut-on à partir de tests effectués sur des échantillons valider ou infirmer ce taux fourni par le constructeur ?

c. On cherche bien sûr à effectuer ce type de test destructif sur le plus petit possible de pièces. Combien de pièces au minimum doivent se trouver dans les échantillons prélevés (et surtout quelle est la confiance que l'on peut alors accorder aux résultats des tests)

6. **L'affaire Castaneda contre Partida**

L'ensemble des faits évoqués ci-dessous est réel.

En Novembre 1976 dans un comté du sud du Texas, Rodrigo Partida était condamné à huit ans de prison pour cambriolage d'une résidence et tentative de viol.
Il attaqua ce jugement au motif que la désignation des jurés de ce comté était discriminante à l'égard des Américains d'origine mexicaine. Alors que 79,1% de la population du comté était d'origine mexicaine, sur les 870 personnes convoquées pour être jurés lors d'une certaine période de référence, il n'y eût que 339 personnes d'origine mexicaine.

Le tirage des jurés s'effectuant "au hasard" dans la population, est-il impossible que, sur 870 personnes convoquées, 870 soient d'origine mexicaine ? 870 ne soient pas d'origine mexicaine ?

II - Echantillonnage

1) Théorèmes d'approximation

Théorème Loi faible des grands nombres

Soit X_1, X_2, ..., X_n, n variables aléatoires indépendantes, de même loi, définies sur Ω, et telles que $E(X_i) = m$ et $V(X_i) = \sigma^2$.
On définit les variables aléatoires :

$$S_n = X_1 + X_2 + \cdots + X_n \quad \text{et} \quad \overline{X_n} = \frac{1}{n} S_n \,.$$

Alors, pour tout $\varepsilon > 0$, $P\left(\left|\overline{X_n} - m\right|\right) < \varepsilon$ tend vers 1 lorsque n tend vers l'infini.
Autrement dit, $\lim\limits_{n\to+\infty} P\left(\left|\overline{X_n} - m\right|\right) = 1$ ou $\overline{X_n}$ converge en probabilité vers m.

Ce théorème fait le lien entre les statistiques et les probabilités, par le fait que, lorsque le nombre d'expériences est (très) grand, on peut choisir comme probabilité d'un événement sa fréquence statistique d'apparition.

Théorème Théorème central limite

Soit X_1, X_2, ...,X_n, n variables aléatoires, indépendantes, de même loi, définies sur Ω, et telles que $E(X_i) = m$ et $V(X_i) = \sigma^2$.

Pour n suffisamment grand, la variable aléatoire $\overline{X_n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ suit approximativement la loi normale $\mathcal{N}\left(m\,;\,\dfrac{\sigma}{\sqrt{n}}\right)$.

Remarque 1 : Si les X_i suivent toutes la même loi normale $\mathcal{N}(m;\sigma)$, alors $\overline{X_n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ suit la loi normale $\mathcal{N}\left(m\,;\,\dfrac{\sigma}{\sqrt{n}}\right)$, et ceci quelque soit n.

L'intérêt fondamental de ce théorème est que, si les X_i suivent une loi quelconque (sans même qu'il soit nécessaire de cette loi, pourvu que ce soit la même pour tous les X_i), alors pour n suffisamment grand, $\overline{X_n}$ suit approximativement la loi $\mathcal{N}\left(m\,;\,\dfrac{\sigma}{\sqrt{n}}\right)$.

Ce théorème justifie l'importance de la loi normale.

Remarque 2 : Pour n suffisamment grand, on sait qu'on peut remplacer les probabilités associées à la loi binomiale $\mathcal{B}(n;p)$ par celles de la loi normale $\mathcal{N}(np;\sqrt{npq})$.

Ceci est un exemple d'application du théorème central limite. En effet, Si X suit la loi $\mathcal{B}(n;p)$, alors X est la somme de n variables de Bernouilli suivant toute la même loi de moyenne p et d'écart type $\sqrt{pq}$.

Alors, d'après le théorème central limite, $\dfrac{X}{n} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ suit approximativement, pour n grand, la loi normale $\mathcal{N}\left(p;\dfrac{\sqrt{pq}}{\sqrt{n}}\right)$, ce qui équivaut à dire que la variable $X = X_1 + X_2 + \cdots + X_n$ suit approximativement la loi normale $\mathcal{N}(np;\sqrt{npq})$.

2) Distribution d'échantillonnage

L'échantillonnage consiste à déterminer les propriétés des échantillons (aléatoires) d'une population, connaissant les propriétés de la population complète.

On considère une population d'effectif N de moyenne m et d'écart type σ.
On prélève avec remise un échantillon aléatoire de taille n.
L'observation de l'échantillon correspond à l'observation de n variables aléatoires X_1 ; X_2 ; ... ; X_n de même loi et de moyenne m et d'écart type σ.

D'après le théorème central limite, la variable aléatoire $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ suit approximativement, pour n suffisamment grand, la loi normale $\mathcal{N}\left(m;\dfrac{\sigma}{\sqrt{n}}\right)$.

Remarques : • Si la taille N de la population est grande, on peut assimiler des tirages successifs sans remise comme des tirages indépendants, la population globale ne changeant prseque pas apès avoir tirer un élément. Dans le cas où N est grand, les tirages sans remise peuvent donc être considérés comme indépendants.

- Les échantillons sont prélevés aléatoirement ; c'est ce qu'on appelle un échantillon représentatif.

Théorème *Soit une population d'effectif N de moyenne m et d'écart type σ. Soit $\overline{X}$ la variable aléatoire qui associe à chaque échantillon de taille n sa moyenne.*
Alors, pour n suffisamment grand, la loi de $\overline{X}$ peut-être approchée par la loi normale $\mathcal{N}\left(m; \dfrac{\sigma}{\sqrt{n}}\right)$.

Exercice 1 Une usine produisant 10 000 objets est réglée pour un poids moyen de 250g, avec un écart type de 10g.
On prélève 200 objets (tirage assimilé à tirage avec remise).
L'échantillon étant suffisamment grand, la loi d'échantillonnage $\overline{X}$ peut-être approchée par la loi normale de moyenne 250 et d'écart type $\dfrac{10}{\sqrt{200}} = \dfrac{\sqrt{2}}{2}$.
Calculer la probabilité pour que la moyenne de l'échantillon soit comprise entre 249g et 251g.

Corollaire *Soit une population d'effetif N dont N' éléments possèdent le caractère étudié. La fréquance du caractère étudié est $p = \dfrac{N'}{N}$.*
Soit la variable aléatoire F donnant la fréquence du caractère étusié pour chaque échantillon aléatoire de taille n prélevé.
Alors, pour n suffisamment grand, la loi de F peut-être approchée par la loi normale $\mathcal{N}\left(p; \sqrt{\dfrac{p(1-p)}{n}}\right)$.

Remarque : Ce corollaire découle directment du théorème précédent, en approchant ici la loi binomiale par une loi normale.

Exercice 2 Au cours d'une consultation électorale, le candidat A a recueilli 55% des voix.
Calculons la probabilité d'avoir, dans un échantillon de taille 100 prélevé parmi les suffrages exprimés, moins de 50% des voix pour le candidat A.
La taille de l'échantillon étant suffisamment grande, F suit approximativement la loi $\mathcal{N}(m;\sigma)$, avec $m = 0,55$ et $\sigma = \sqrt{\dfrac{p(1-p)}{n}} \simeq 0,05$.
Calculer la probabilité d'avoir moins de 50% des voix dans un échantillon de taille 100.

III - Statistiques inférentielles : estimation

IV - Statistiques inférentielles : tests de validité d'hypothèse