Modélisation statistique - TD1
Echantillonage aléatoire simple

On s'intéresse à une population finie de taille N, et à un échantillon de taille n. On note u_α, $\alpha = 1, \cdots, N$ les individus de la population et u_{α_i}, $i = 1, \cdots, n$ les individus de l'échantillon. On considère X un caractère quantitatif, x_α est alors la mesure X faite sur l'individu u_α. On considère

— μ la moyenne de la population pour le caractère X

$$\mu = \frac{1}{N} \sum_{\alpha=1}^{N} x_\alpha,$$

— σ^* l'écart type corrigé de la population

$$\sqrt{\frac{1}{N-1} \sum_{\alpha=1}^{N} (x_\alpha - \mu)^2}.$$

L'étude du caractère X dans la population porte généralement au minimum sur ces deux dernières valeurs, la moyenne et l'écart-type du caractère dans la population. Comment construire un échantillon pour obtenir une bonne approximation de ces paramètres, tout en contrôlant l'erreur d'échantillonnage ?

On appelle ***plan aléatoire simple*** *un échantillon obtenu par une méthode qui assure à chaque échantillon possible la même probabilité d'être sélectionné. On appelle* ***échantillon aléatoire simple*** *un échantillon obtenu par un plan aléatoire simple.*

Il existe deux types de plans aléatoires simples, selon qu'il y a ou non remises. Dans ce TD, on s'intéressera uniquement à l'échantillonnage aléatoire simple **sans remise**. On définit sur l'ensemble des échantillons possibles les variables aléatoires

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_{\alpha_i} \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{\alpha_i} - \bar{Y})^2.$$

Ces variables aléatoires sont aléatoires, car le choix des individus de l'échantillon α_i est aléatoire. En effet x_{α_i} n'est pas aléatoire, on suppose qu'il n'y a pas d'erreur de mesures. Ces variables sont des estimateurs de μ et σ^{*2}.

Propriété 1

$$E(\bar{X}) = \mu \qquad \mathbf{E}(\bar{X}) = \mu$$

où **E** est l'espérance sur tous les échantillons possibles. Cette première propriété signifie que la moyenne des moyennes sur tous les échantillons possibles est égale à la moyenne de la population. Le plan aléatoire simple ne commet pas d'erreur systématique, il est sans biais. Cependant, la portée de cette propriété reste limitée, car elle concerne la moyenne sur tous les échantillons et ne nous dit rien sur un échantillon particulier.

Propriété 2

$$V(\bar{X}) = \frac{\sigma^{*2}}{n}(1-f)$$

avec $V(\bar{X})$ la variance des moyennes sur tous les échantillons possibles, et $f = \frac{n}{N}$ le taux de sondage. σ^* est estimé par S. Cette seconde propriété nous renseigne sur l'erreur d'échantillonnage $\bar{X} - \mu$, en effet $V(\bar{X} - \mu) = V(\bar{X})$. La racine de $V(\bar{X})$ est donc une mesure de l'erreur de mesure.
On a

$$\sqrt{V(\bar{X})} = \frac{\sigma^*}{\sqrt{n}}\sqrt{1-f}$$

On apprend que plus la taille de l'échantillon augmente, plus l'erreur d'échantillonnage diminue. Cependant, la diminution n'est pas proportionnelle à la taille de l'échantillon, et la multiplication de la taille de l'échantillon par 4, ne fera que diminuer l'erreur par 2.

Propriété 3

$$\bar{X} \underset{\substack{n\to\infty \\ N\to\infty \\ n/N=f}}{\longrightarrow} \mathcal{N}(\mu, \frac{\sigma^{2*}}{n}(1-f))$$

La loi de $\bar{X}$ tend vers une loi normale $\mathcal{N}(\mu, \frac{\sigma^{2*}}{n}(1-f))$ lorsque n et N tendent vers l'infini et le rapport $\frac{n}{N}$ est constant. Cette propriété nous informe sur la loi de l'erreur de mesure $\bar{X} - \mu$. Pour 95% des échantillons, l'erreur de mesure est comprise dans l'intervalle $[-1.96\frac{\sigma^*}{\sqrt{n}}\sqrt{1-f}; 1.96\frac{\sigma^*}{\sqrt{n}}\sqrt{1-f}]$. Mais attention, cette propriété ne nous dit rien pour un échantillon en particulier. Les propriétés 1 et 2 seront démontrées en TD.

1 Exercice : Mise en application de l'échantillonnage aléatoire simple

Des cercles sont représentés sur la fiche "population d'objets". On souhaite estimer la moyenne des aires de ces formes.

1. Construisez un échantillon de 12 cercles par une méthode quelconque (lancer de crayons, choix raisonné, formes ayant une intersection avec une droite quelconque,...). Noter les numéros des objets choisis.
2. Les aires de chacun des cercles sont notés sur la table `PopAire.csv`. A l'aide du logiciel **R** calculer l'aire moyenne de votre échantillon. Calculer ensuite la moyenne et l'écart-type corrigé de la population.
3. Calculer la moyenne et l'écart-type des estimations obtenues dans le groupe de TD. Construire l'histogramme des estimations obtenues dans le groupe de TD. Que peut-on

dire sur la méthode d'échantillonnage "au hasard" mise en place ?

4. Reprendre les questions 1 à 3, en réalisant cette fois un échantillonnage aléatoire simple.

2 Exercice : Démonstration des résultats de l'échantillonnage aléatoire simple

On se place dans le cadre de l'échantillonnage aléatoire simple. On note n la taille de l'échantillon et ϵ_i la variable aléatoire définie par

$$\begin{cases} 1 \text{ si } u_i \in \text{l'échantillon} \\ 0 \text{ sinon,} \end{cases}$$

où u_i est l'unité i de la population.

1. Démontrer que $\mathbf{P}(\epsilon_i = 1) = \frac{n}{N}$ et que $\mathbf{P}((\epsilon_i = 1) \cap (\epsilon_j = 1)) = \frac{n(n-1)}{N(N-1)}$ si $i \neq j$. En déduire que $\mathbf{E}(\epsilon_i) = \frac{n}{N}$ et que $\mathbf{E}((\epsilon_i \epsilon_j) = \frac{n(n-1)}{N(N-1)}$ si $i \neq j$.
2. En déduire $\mathbf{V}(\epsilon_i)$ et $\mathbf{Cov}(\epsilon_i, \epsilon_j)$.
3. Soit $\overline{X}$ la moyenne de l'échantillon, démontrer que $\overline{X} = \frac{1}{n} \sum_{\alpha=1}^{N} x_\alpha \epsilon_\alpha$. En déduire que $\mathbf{E}(\overline{X}) = \mu$.
4. Démontrer la formule $\mathbf{V}(\overline{X}) = \frac{\sigma^{*2}}{n}(1 - \frac{n}{N})$.