Echantillonnage - Estimation

1 Echantillonnage

On suppose **connue** une population $\mathcal{P}$ de taille N, de moyenne m et d'écart-type σ. On note p la proportion d'individus de $\mathcal{P}$ vérifiant la propriété A. Tous les échantillons extraits de $\mathcal{P}$ seront supposés aléatoires et non exhaustifs (tirages assimilés à des tirages avec remise).

Théorème 1 Soit X la variable aléatoire qui, à un individu de $\mathcal{P}$ choisi au hasard, associe la valeur de cet individu. Alors : $E(X) = m$ et $\sigma(X) = \sigma$

Démonstration. Soit la population $\mathcal{P}$:

Valeur	x_1	x_2	$\cdots$	x_p
Effectif	n_1	n_2	$\cdots$	n_p

.

$N = \sum_{k=1}^{p} n_k$

$P(X = x_k) = \frac{n_k}{N} \Rightarrow E(X) = \sum x_k P(X = x_k) = \sum x_k \frac{n_k}{N} = \frac{1}{N} \sum n_k x_k = m.$

$\sigma^2(X) = E\left(X^2\right) - E(X)^2 = \sum x_k^2 P(X = x_k) - m^2 = \sum x_k^2 \frac{n_k}{N} - m^2 = \frac{1}{N} \sum n_k x_k^2 - m^2 = \sigma^2.$

Remarque 1 Si X suit une loi normale, on dit que la distribution de $\mathcal{P}$ est normale.

Soit $\mathcal{E}$ un échantillon de taille n. On notera :
- X_1 la variable aléatoire qui, au 1^{er} individu de $\mathcal{P}$ choisi au hasard, associe sa valeur.
- X_2 la variable aléatoire qui, au $2^{ème}$ individu de $\mathcal{P}$ choisi au hasard, associe sa valeur.
- .....
- X_n la variable aléatoire qui, au $n^{ème}$ individu de $\mathcal{P}$ choisi au hasard, associe sa valeur.

Remarque 2 Les échantillons étant aléatoires et non exhaustifs, les variables aléatoires $X_1, X_2, \ldots, X_n$ sont indépendantes et de même loi que X.

1.1 Moyenne d'échantillon.

Théorème 2 La variable aléatoire $\bar{X}_n$ qui, à un échantillon de taille n, associe sa moyenne, vérifie :

$E(\bar{X}_n) = m$	et	$\sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$

De plus, d'après le théorème central limite, pour n assez grand ($n \geq 30$), $\bar{X}_n \underset{\sim}{\hookrightarrow} \mathcal{N}(m; \frac{\sigma}{\sqrt{n}})$

Remarque 3 $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$

Remarque 4 La loi suivie par $\bar{X}_n$ est inconnue. On sait seulement que c'est *approximativement* une loi normale dans le cas de grands échantillons ($n \geq 30$).

Par contre, si $X \hookrightarrow \mathcal{N}(m; \sigma)$, alors $\bar{X}_n \hookrightarrow \mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$. *Il faut le montrer !!!* Les remarques **??** et **??** sont alors fondamentales.

1.2 Variance d'échantillon.

Théorème 3 La variable aléatoire $S^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$ qui, à un échantillon de taille n, associe sa variance, vérifie : $\boxed{E(S^2) = \frac{n-1}{n}\sigma^2}$

1.3 Fréquence d'échantillon.

Théorème 4 La variable aléatoire F qui, à un échantillon de taille n, associe la proportion d'individus de cet échantillon vérifiant la propriété A vérifie $nF \hookrightarrow \mathcal{B}(n;p)$

Corollaire 1 $\boxed{E(F) = p \quad \text{et} \quad \sigma(F) = \sqrt{\frac{p(1-p)}{n}}}$

Remarque 5 Pour n assez grand ($n \geq 30$), F suit approximativement la loi $\mathcal{N}\left(p; \sqrt{\frac{p(1-p)}{n}}\right)$ d'après le théorème central limite.

2 Estimation

2.1 Estimation ponctuelle.

Une estimation ponctuelle (sans biais) d'un paramètre **inconnu** θ de la population $\mathcal{P}$ est la valeur $\widehat{\theta}$ prise sur un échantillon $\mathcal{E}$ par une variable aléatoire $\widetilde{\theta}$ qui vérifie $E\left(\widetilde{\theta}\right) = \theta$.
On dit alors que $\widetilde{\theta}$ est une estimateur de θ.

2.1.1 Estimation ponctuelle d'une moyenne.

La variable aléatoire $\bar{X}_n$ étudiée lors de l'échantillonnage vérifie $E\left(\bar{X}_n\right) = m$.
$\bar{X}_n$ est donc un estimateur de m.
Sur un échantillon $\mathcal{E}$, $\bar{X}_n$ prend pour valeur la moyenne de cet échantillon m_e

Théorème 5 La moyenne m_e d'un échantillon aléatoire d'une population $\mathcal{P}$ est une estimation ponctuelle $\widehat{m}$ de la moyenne m de cette population. $\boxed{\widehat{m} = m_e}$

2.1.2 Estimation ponctuelle d'une variance.

La variable aléatoire S^2 étudiée lors de l'échantillonnage vérifie $E\left(S^2\right) = \frac{n-1}{n}\sigma^2$.
Donc la variable aléatoire $\widetilde{S^2} = \frac{n}{n-1}S^2$ vérifie $E\left(\widetilde{S^2}\right) = \sigma^2$.
$\widetilde{S^2}$ est donc un estimateur de σ^2.
Sur un échantillon $\mathcal{E}$, S^2 prend pour valeur la variance de cet échantillon σ_e^2.

Théorème 6 Si σ_e^2 est la variance d'un échantillon aléatoire d'une population $\mathcal{P}$, $\frac{n}{n-1}\sigma_e^2$ est une estimation ponctuelle $\widehat{\sigma^2}$ de la variance inconnue σ^2 de cette population. $\boxed{\widehat{\sigma^2} = \frac{n}{n-1}\sigma_e^2}$

Remarque 6 On prendra $\widehat{\sigma} = \sigma_{n-1} = \sqrt{\frac{n}{n-1}\sigma_e^2} = \sqrt{\frac{n}{n-1}}\sigma_e$ comme estimation ponctuelle de σ.

2.1.3 Estimation ponctuelle d'une proportion.

La variable aléatoire F étudiée lors de l'échantillonnage vérifie $E(F) = p$.
F est donc un estimateur de p.
Sur un échantillon $\mathcal{E}$, F prend pour valeur la proportion p_e d'individus de cet échantillon vérifiant une propriété A.

Théorème 7 La proportion p_e d'individus vérifiant une propriété A dans un échantillon aléatoire d'une population $\mathcal{P}$ est une estimation ponctuelle $\widehat{p}$ de la proportion inconnue p des individus de $\mathcal{P}$ vérifiant A.

$$\boxed{\widehat{p} = p_e}$$

2.1.4 Estimation ponctuelle de $\sigma(F) = \sqrt{\frac{p(1-p)}{n}}$

Théorème 8 Si p_e est la proportion d'individus vérifiant une propriété A dans un échantillon aléatoire d'une population $\mathcal{P}$ alors $\frac{p_e(1-p_e)}{n-1}$ est une estimation ponctuelle de $\frac{p(1-p)}{n}$

Remarque 7 On prendra $\sqrt{\frac{p_e(1-p_e)}{n-1}}$ comme estimation ponctuelle de $\sigma(F) = \sqrt{\frac{p(1-p)}{n}}$

2.2 Estimation par intervalle de confiance

En conservant les notations de (**??**), un intervalle de confiance de θ au niveau de confiance c est la valeur prise sur un échantillon $\mathcal{E}$ par l'intervalle aléatoire $\left[\widetilde{\theta} - h; \widetilde{\theta} + h\right]$ tel que

$$P\left(\theta \in \left[\widetilde{\theta} - h; \widetilde{\theta} + h\right]\right) = c$$

2.2.1 Estimation par intervalle d'une moyenne

On supposera que $\bar{X}_n$ suit une loi normale (ce qui est réalisé si la distribution de $\mathcal{P}$ est normale)
Il faut déterminer h tel que $P(m \in [\bar{X}_n - h; \bar{X}_n + h]) = c$.
La valeur prise sur un échantillon $\mathcal{E}$ par $[\bar{X}_n - h; \bar{X}_n + h]$ est l'intervalle cherché.
Si on pose $c = 2\Pi(\alpha) - 1$, on trouve (faire les calculs) : $\left[m_e - \alpha\frac{\sigma}{\sqrt{n}}; m_e + \alpha\frac{\sigma}{\sqrt{n}}\right]$.

Théorème 9 Soit $\mathcal{P}$ une population de moyenne inconnue m et d'écart-type σ.
L'intervalle $\left[m_e - \alpha\frac{\sigma}{\sqrt{n}}; m_e + \alpha\frac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de m au niveau de confiance $2\pi(\alpha) - 1$.

Remarque 8 Si σ est inconnu on utilisera la remarque **??**

2.2.2 Estimation par intervalle d'une fréquence

On supposera que F suit une loi normale.
Il faut déterminer h tel que $P(p \in [F - h; F + h]) = c$.
La valeur prise sur un échantillon par $[F - h; F + h]$ est l'intervalle cherché.
Si on pose $c = 2\Pi(\alpha) - 1$, on trouve $\left[p_e - \alpha\sqrt{\frac{p_e(1-p_e)}{n-1}}; p_e + \alpha\sqrt{\frac{p_e(1-p_e)}{n-1}}\right]$ (utilisation de la remarque **??**)

Théorème 10 Soit $\mathcal{P}$ une population ayant une proportion inconnue p d'individus vérifiant la propriété A. L'intervalle $\left[p_e - \alpha\sqrt{\frac{p_e(1-p_e)}{n-1}}; p_e + \alpha\sqrt{\frac{p_e(1-p_e)}{n-1}}\right]$ est un intervalle de confiance de p au niveau de confiance $2\pi(\alpha) - 1$.