Estimation

1 Principe de l'estimation

La théorie de l'échantillonnage consistait à déterminer des propriétés sur des échantillons tirés au hasard parmi une population dont on connait les propriétés.
Le principe de l'estimation est de faire exactement l'inverse, c'est-à-dire que l'on a accès à des informations sur des échantillons (sondages, tests de conformité,...) et l'on souhaite déterminer certaines propriétés sur la population entière.

Remarque 1. Il est clair que l'on ne pourra jamais obtenir, à partir d'un échantillon réduit, des données exactes sur la population entière, c'est pourquoi il sera important dans la suite de donner des estimations de certaines données mais en précisant **toujours** la marge d'erreur ou le risque que l'on prend.

Exemple 1. Si vous connaissez les notes au BAC de trois de vos amis, il est assez douteux d'en déduire que la moyenne de leurs notes est proche de la moyenne du BAC de l'année où ils l'ont passé. Par contre, si vous connaissez les notes de 50 amis, il est probable que la moyenne de leur notes sera plus proche de la moyenne du BAC que celle de vos trois amis. En effet, vous avez un échantillon beaucoup plus grand ce qui vous permet d'être a priori mieux informé et de moins tenir compte des cas particuliers.

On considérera dans la suite uniquement des tirages aléatoires d'échantillons. Le tirage d'éléments dans une population peut-être fait de façon exhaustive (c'est-à-dire sans remise) ou de façon non-exhaustive (avec remise). Dans ce dernier cas, les tirages sont indépendants.

En pratique, lorsque la population a un grand effectif, on tire seulement un faible nombre d'éléments et l'on assimile un tirage sans remise à un tirage avec remise.

2 Estimation d'une moyenne

Soit X une variable aléatoire définie sur la population mère Ω de taille N et $\mathcal{E}$ un échantillon de taille $n \leq N$ issu de cette population. On suppose que l'on a calculé la moyenne μ_e et l'écart-type σ_e de cet échantillon. Notre but est d'approcher au mieux la moyenne μ et l'écart-type σ de la population.

La proposition suivante propose une première estimation.

Proposition 1. Une estimation ponctuelle $\hat{\mu}$ de la moyenne μ de la population est:

$$\hat{\mu} = \mu_e.$$

Une estimation ponctuelle $\hat{\sigma}$ de l'écart-type σ_e de la population est donné par:

$$\hat{\sigma} = \sqrt{\frac{n}{n-1}}\sigma_e.$$

Le coefficient $\sqrt{\frac{n}{n-1}}$ peut sembler étrange, il s'appelle le correcteur de biais. Il apparait car comme la moyenne μ de la population est inconnue, on utilise la moyenne de l'échantillon à sa place. Le correcteur de biais permet alors en général de rapprocher la valeur de $\hat{\sigma}$ de la vraie valeur σ recherchée.

3 Estimation d'une moyenne par intervalle de confiance

Maintenant, on veut obtenir des résultats plus précis qu'avant. En fait, on veut pouvoir estimer notre risque d'erreur.

Nous avons vu en faisant de l'échantillonnage que la moyenne d'un échantillon de taille $n \geq 30$ suit approximativement une loi normale: $\overline{X} \hookrightarrow \mathcal{N}\left(\mu; \frac{\sigma^2}{n}\right)$.

On cherche un intervalle qui contient avec une confiance arbitraire C (95% ou 99% par exemple) la moyenne μ de la population. On cherche donc un rayon r tel que:

$$\mathbb{P}(\overline{X} - r \leq \mu \leq \overline{X} + r) = C.$$

En faisant un petit calcul (que l'on a déjà fait plusieurs fois dans les exercices), on s'aperçoit que

$$\mathbb{P}(\overline{X} - r \leq \mu \leq \overline{X} + r) = \mathbb{P}(\mu - r \leq \overline{X} \leq \mu + r)$$

Ainsi, on cherche un rayon r tel que:

$$\mathbb{P}(\mu - r \leq \overline{X} \leq \mu + r) = C.$$

Pour trouver cette valeur de r, on se ramène à une variable aléatoire de loi normale centrée réduite:

$$\mathbb{P}(\mu - r \leq \overline{X} \leq \mu + r) = \mathbb{P}\left(\frac{\mu - r - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\mu + r}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(\frac{-r\sqrt{n}}{\sigma} \leq \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{r\sqrt{n}}{\sigma}\right).$$

On obtient alors après quelques calculs:

$$\mathbb{P}(\mu - r \leq \overline{X} \leq \mu + r) = 2\Pi\left(\frac{r\sqrt{n}}{\sigma}\right) - 1.$$

Il suffit donc de lire sur la table de la loi normale pour quelle valeur de $\frac{r\sqrt{n}}{\sigma}$ est-ce que:

$$\Pi\left(\frac{r\sqrt{n}}{\sigma}\right) = \frac{C+1}{2}.$$

Si l'on note t cette valeur, alors on obtient la formule:

$$\boxed{r = t\frac{\sigma}{\sqrt{n}}.}$$

Exemple 2.
Si l'on choisit une confiance de 95%, alors on cherche t tel que $\Pi(t) = \frac{0,95+1}{2} = 0,975$.
On obtient alors $t = 1,96$.
Avec une confiance de 99%, on aurait obtenu $t = 2,575$.

A présent, si vous avez des informations collectées sur un échantillon de taille n et que vous avez calculé μ_e, on déduit que l'intervalle de confiance à $C\%$ est l'intervalle:

$$\left[\mu_e - t\frac{\sigma}{\sqrt{n}}; \mu_e + t\frac{\sigma}{\sqrt{n}}\right].$$

Il suffit alors de remplacer par les valeurs connues.

Remarque 2.

- Si l'écart-type de la population totale σ n'est pas connu, alors on le remplace par son estimation ponctuelle: $\hat{\sigma}$.
- Plus le degré de confiance C est grand, plus l'intervalle de confiance sera étendu.

Exemple 3. Une université comporte 1500 étudiants. On mesure la taille de 40 d'entre eux. La moyenne μ_e et l'écart-type σ_e calculés à partir de cet échantillon sont:

$$\mu_e = 176cm \text{ et } \sigma_e = 6cm.$$

Trouver une estimation ponctuelle des paramètres de la population puis déterminer une estimation de μ par un intervalle de confiance à 95%.

4 Estimation d'une proportion

On peut bien sur faire la même chose avec des proportions. On considère donc un population qui possède (ou pas) un certain caractère. On ne connait pas la proportion p de la population qui possède le caractère et l'on cherche à l'estimer grâce à un échantillon de taille n où l'on a mesuré une proportion p_e d'individu avec ce caractère.

Proposition 2. Une estimation ponctuelle $\hat{p}$ de la proportion p d'individus avec le caractère dans la population est:

$$\hat{p} = p_e.$$

Une estimation ponctuelle $\hat{\sigma_p}$ de l'écart-type σ_p est donné par:

$$\begin{cases} \sqrt{\frac{p_e(1-p_e)}{n-1}} & \text{si } n \leq 30 \\ \sqrt{\frac{p_e(1-p_e)}{n}} & \text{si } n > 30 \end{cases}$$

5 Estimation d'une proportion par intervalle de confiance

Si F est la variable aléatoire correspondant à la proportion d'un caractère dans la population dans un échantillon de taille n, alors on cherche un intervalle qui contient p avec une confiance arbitraire C (95% ou 99%,...).

On cherche donc un rayon r tel que:

$$\mathbb{P}(F - r \leq p \leq F + r) = C.$$

Ceci revient à chercher r tel que:

$$\mathbb{P}(p - r \leq F \leq p + r) = C.$$

Mais comme $F \hookrightarrow \mathcal{N}(p; \sigma_p)$, on cherche r tel que:

$$\mathbb{P}\left(\frac{-r}{\sigma_p} \leq \frac{F-p}{\sigma_p} \leq \frac{r}{\sigma_p}\right) = C.$$

Et donc

$$2\Pi\left(\frac{r}{\sigma_p}\right) - 1 = C.$$

Il suffit donc de lire sur la table de la loi normale pour quelle valeur de $\frac{r}{\sigma_p}$ est-ce que:

$$\Pi\left(\frac{r}{\sigma_p}\right) = \frac{C+1}{2}.$$

Si l'on note t cette valeur, on obtient la formule:

$$\boxed{r = t\sigma_p.}$$

A présent, si vous avez des informations collectées sur un échantillon de taille n et que vous avez calculé p_e, on déduit que l'intervalle de confiance à $C\%$ est l'intervalle:

$$\left[p_e - t\sigma_p; p_e + t\sigma_p\right].$$

Il suffit alors de remplacer par les valeurs connues.

Exemple 4. A quelques jours d'une élection, un candidat fait faire un sondage. Sur les 150 personnes interrogées, 45 se disent prêtes à voter pour lui aux prochaines élections.
Déterminer une estimation ponctuelle des paramètres puis une estimation de p par intervalle de confiance à 80%.