Chapitre 4

Au-delà de l'échantillon : Estimation des paramètres d'une population

Nous supposons que vous disposez d'un **échantillon aléatoire simple** d'un nombre n d'individus issus de la population parent dont on souhaite maintenant estimer les paramètres.

♠ **Attention** : Tout autre mode d'échantillonnage vous empêche d'estimer ces paramètres. Nul besoin dans ce cas de lire ce chapitre...

4.1 estimation, estimateur et biais

Revenons dans la forêt du Gâvre et imaginons que vous vouliez estimer la **proportion** p de conifères dont le diamètre pris à un mètre du sol est supérieur à 40 cm ainsi que le diamètre **moyen** de l'ensemble des conifères de la forêt, noté μ. Vous ne pouvez évidemment pas mesurer tous les arbres et vous avez effectué un échantillonnage aléatoire simple de 50 individus, nombre qui vous a semblé suffisant pour espérer que la proportion $\hat{p}$ et la moyenne $\bar{x}$ obtenues fournissent de bonnes estimations.

Pourquoi ? L'argument repose parfois sur l'intuition mais on subodore ici l'utilité du chapitre « Théorèmes limites » au programme de deuxième année... Notre objectif est de voir comment l'exploiter dans les deux cas évoqués dans l'exemple ci-dessu : estimation d'une proportion ou estimation d'une moyenne.

♠ **Attention** : L'ensemble de ce qui va suivre n'a de sens que si vos données sont non **biaisées**, c'est-à-dire ne souffrent pas d'une erreur systématique, liée à l'expérimentateur, aux conditions d'expériences ou encore à l'appareil de mesure..

Définition : Si X est une variable aléatoire dont la loi dépend d'un paramètre θ (par exemple $\theta = m$ ou $\theta = \sigma$ dans le cas de la loi normale) et $(X_1, ..., X_n)$ un n-échantillon de X. Une variable aléatoire T_n, fonction de $(X_1, ..., X_n)$ est un **estimateur sans biais** et convergent de θ si $\mathbb{E}(T_n) = \theta$ et $\lim\limits_{n \to +\infty} \mathbb{V}(T_n) = 0$

Autrement dit l'estimateur est sans biais s'il permet de fournir une valeur approchée du paramètre cherché d'autant plus proche que le nombre de valeurs de l'échantillon est important et ce sans le sous-estimer ou le surestimer de façon systématique.

Plusieurs estimateurs sont possibles et T'_n sera dit **meilleur estimateur** que T_n si $\mathbb{V}(T'_n) \leq \mathbb{V}(T_n)$.

Les estimateurs courants sont :

– la moyenne $\bar{m}$ de l'échantillon pour estimer la moyenne μ de la population.

– le rapport $\dfrac{k}{n}$ pour estimer une proportion.

– l'écart-type $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{m})^2}$ pour estimer l'écart-type σ de la population.

4.2 estimation d'une proportion par intervalle

Nous avons à notre disposition un échantillon aléatoire simple de n individus et nous souhaitons inférer à partir de ses seules valeurs la proporion p d'un caractère donné au sein de la population dont il est issu. Les exemples sont nombreux : proportion d'une essence donnée dans une forêt, proportion de grains d'un diamètre inférieur à $2mm$ dans un sable, proportion de micas dans un granit, etc...

Commençons par formaliser le contexte dans lequel nous travaillons :

– Soit Ω la population au sein de laquelle on effectue un tirage et soit X la variable aléatoire de Bernoulli égale à 1 si l'individu extrait possède le caractère étudié.

– la suite $\{X_1,\ X_2,\ ...,\ X_n\}$ de variables de même loi que X modélise le résultat de notre échantillonnage aléatoire simple.

– Soit $x = X_1 + ... + X_n$. Alors, sous réserve que notre échantillon soit correctement élaboré (indépendance, propabilité du succès invariable au cours des prélèvements), on a $x \hookrightarrow \mathcal{B}(n, p)$ puisque x dénombre les succès au cours de n épreuves indépendantes de Bernoulli de même probabilité du succès p.

– On suppose par ailleurs que $\boxed{n > 30}$ ou bien $\boxed{n \cdot p \geq 10}$ ou bien encore $\boxed{npq \geq 5}$ (ces conditions sont plus ou moins strictes selon les ouvrages.). On peut alors approximer la loi de binomiale par la loi normale de paramètres $m = np$ et $s = \sqrt{npq}$.

– Posons pour terminer $\hat{p} = x/n$ pour désigner la proportion d'individus qui, au sein de l'échantillon, possèdent le caractère étudié.

Au regard des hypothèses précédentes qu'il faudra justifier, nous admettrons désormais que nous avons approximativement $\hat{p} = \dfrac{x}{n} \hookrightarrow \mathcal{N}(p, \sqrt{\dfrac{pq}{n}})$

conséquences En centrant et réduisant, on obtient pour tout a réel positif :

$$\mathbb{P}\left[p - a\sqrt{\frac{pq}{n}} < \hat{p} < p + a\sqrt{\frac{pq}{n}}\right] = \mathbb{P}(-a < \hat{p}^* < a) \text{ où } \hat{p}^* \hookrightarrow \mathcal{N}(0,1)$$

Conformément à l'ensemble des ouvrages de statistique, nous poserons désormais $a = z_{\alpha/2}$ et nous remplaçons le produit pq par $\hat{p}\hat{q}$ puisque p n'est pas encore connu. Dès lors :

$$\mathbb{P}\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right] = \mathbb{P}(-z_{\alpha/2} < \hat{p}^* < z_{\alpha/2}) = 2\phi(z_{\alpha/2}) - 1$$

Voici les probabilités obtenues pour des valeurs usuelles de $z_{\alpha/2}$ (On notera $E = z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$ qu'on appellera ***marge d'erreur*** pour les proportions, parfois encore ***erreur maximale de l'estimation***) :

- si $z_{\alpha/2} = 1,645$, $\mathbb{P}(\hat{p} - E < p < \hat{p} + E) = 2\phi(1,645) - 1 = 0,9$
- si$z_{\alpha/2} = 1,96$, $\mathbb{P}(\hat{p} - E < p < \hat{p} + E) = 2\phi(1,96) - 1 = 0,95 = 1 - 0.5$

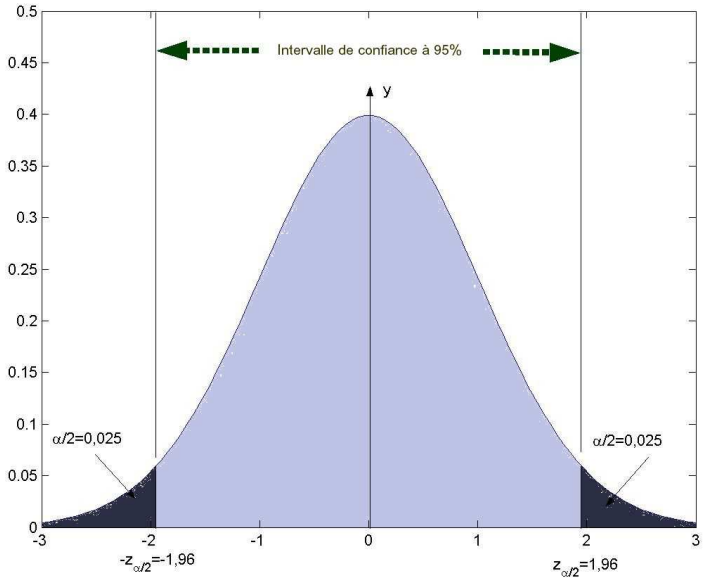

FIG. 4.1 – courbe de Gauss centrée réduite

Définition : L'intervalle $IC = [\hat{p} - E; \hat{p} + E]$ tel que $\mathbb{P}(\hat{p} - E < p < \hat{p} + E) = 1 - \alpha$ est appelé **intervalle de confiance** ou **intervalle de sécurité** au niveau $1 - \alpha$. La probabilité que cet intervalle contienne p vaut $(1 - \alpha)$ qu'on appellera le **seuil de confiance**.

♠ Attention : Ne pas écrire que « la probabilité que p appartienne à l'intervalle de confiance vaut $1-\alpha$ » car p est fixé et n'est pas une variable aléatoire.

Remarque : La probabilité que cet intervalle ne contienne pas p vaut α qu'on appelle également le **seuil de risque**. On peut donc commettre une erreur de probabilité α en assurant que p se trouve entre $\hat{p}-E$ et $\hat{p}+E$.

Mise en pratique :

Exemple 1 : Disposant de la proportion de filles en BCPST2 à l'Externat, je cherche à estimer la proportion de filles inscrites au concours au niveau national. En 2008, il y avait 23 filles sur les 33 étudiants de la promo 14, soit $\hat{p} = 23/33 = 0,697$.
$n \geq 30$ se qui permet d'utiliser l'approximation par la loi normale.
On a : $E = 1,96 \cdot \sqrt{\dfrac{0,697(1-0,697)}{33}} = 0,157$
L'intervalle de confiance au niveau $1-0,5 = 0,95$ est donc :

$$IC = [0,697 - 0,157; 0,697 + 0,157] = [0,54; 0,854]$$

Conclusion : La probabilité que IC ne contienne pas p est de 0.5 ou encore « il y a 95% de chances que IC contienne p ». Ce qu'on vérifie puisque les statistiques nationales pour l'année 2008 assurent que 1913 filles étaient inscrites sur les 2704 candidats. Soit $p = 0,707$

♠ **Attention** : Si je prends 100 classes de BCPST2, je dois m'attendre à ce que 5 d'entre elles fournissent un Intervalle de Confiance IC qui ne contient pas $p = 0,707..$!

Exemple 2 : En analysant un échantillon aléatoire simple de 1236 grains issus d'une surface de sédimentation, on a obtenu que 12% d'entre eux avait son diamètre supérieur à $2mm$.
$n >> 30$ et les conditions d'approximation par la loi normale sont satisfaites.
Au seuil de risque de 5% on a : $E = 1,96 \cdot \sqrt{\dfrac{0,12(1-0,12)}{1236}} = 0,018$
L'intervalle de confiance au niveau $1-0,5 = 0,95$ est donc :

$$IC = [0,12 - 0,018; 0,12 + 0,018] = [0,102; 0,138]$$

Remarque : Si on effectue 20 échantillonnages sur la même surface, on doit s'attendre que l'un d'entre eux, avec un niveau de confiance de 95%, ne contienne pas la valeur p cherchée...

4.3 estimation d'une moyenne par intervalle

Ce qui est en jeu ici est le théorème de la limite centrée, théorème essentiel énoncé dans le chapitre sur les théorèmes limites.

Rappel : Si $(X_i)_{i\in\mathbb{N}^*}$ est une suite de variables aléatoires indépendantes définies sur un même espace probabilisé $(\Omega, \mathcal{P}(\Omega), P)$ qui suivent la même loi d'espérance μ et d'écart-type σ, alors pour tout x réel :

$$\lim_{n\to+\infty} \mathbb{P}(S_n \leq x) = \phi_{(n\mu,\sigma\sqrt{n})}(x) \text{ où } S_n = \sum_{i=1}^{n} X_i.$$

Plus intuitivement, on peut dire que, lorsque n devient suffisamment grand, on peut approcher la loi de S_n par la loi normale $\mathcal{N}(n\mu, \sigma\sqrt{n})$ ou bien la loi de $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ par la loi normale $\mathcal{N}(\mu, \sigma/\sqrt{n})$ ou bien encore la loi de S_n^* par la loi normale centrée réduite.

Or nous sommes dans le cas d'un échantillon aléatoire simple pour lequel vos observations sont indépendantes et peuvent être considérées comme n réalisations de la *loi parente* X.

Conséquence : Si X suit une loi normale de paramètres μ et σ ou bien si vous avez plus de 30 résultats, $\bar{x}$ suit *exactement* ou *approximativement* une loi normale de paramètres μ et $\sigma/\sqrt{n}$.

Justifions-le :

- Si X suit une loi normale de paramètres μ et σ (ce que vous vérifierez en vous contentant d'un test de normalité sur vos échantillons[1]) alors votre moyenne notée $\bar{x}$ suit *exactement* une loi normale de paramètres μ et $\sigma/\sqrt{n}$.
  Rappelons en effet que si $X_1 \hookrightarrow \mathcal{N}(\mu_1, \sigma_1)$ et $X_2 \hookrightarrow \mathcal{N}(\mu_2, \sigma_2)$, X_1 et X_2 étant indépendantes, alors : $X_1 + X_2 \hookrightarrow \mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
  Par récurrence, si X_1,..., X_n est une suite de n variables aléatoires indépendantes de même loi normale de paramètres μ et σ alors $X_1 + ... + X_n \hookrightarrow \mathcal{N}(n \cdot \mu, \sqrt{n\sigma^2})$
  D'où $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \hookrightarrow \mathcal{N}(\frac{n\cdot\mu}{n}, \frac{\sqrt{n\sigma^2}}{n})$ ou encore $\bar{x} \hookrightarrow \mathcal{N}(\mu, \sigma/\sqrt{n})$
- Si la taille de votre échantillon est supérieure à 30, le théorème de la limite centrée assure que c'est une variable aléatoire qui suit *approximativement* une loi normale de paramètres μ et $\sigma/\sqrt{n}$.

[1] distribution « en gros » symétrique, avec un seul mode et sans valeurs extrêmes. Comparaison avec la courbe de Gauss de paramètres $\bar{x}$ et s, droite de Henry, etc.

La question fondamentale est de savoir ce qu'il est possible de dire de la moyenne obtenue sur notre échantillonnage... Est-elle proche de la moyenne μ de la population ? quelle estimation de μ pouvons-nous espérer à opartir de notre expérience ? Quelle précision aura cette estimation ?

Imaginons quelques instants que vous connaissiez σ, hypothèse le plus souvent absurde puisque, ne connaissant pas μ, il y a peu de chances que vous connaissiez σ... M'enfin, voilà ce qu'on pourrait écrire :
Pour tout $z_{\alpha/2}$ réel positif, en centrant et réduisant :

$$\mathbb{P}[\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = \mathbb{P}(z_{\alpha/2} < \bar{x}^* < z_{\alpha/2})$$

Notons comme précédemment $\boxed{E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$ la marge d'erreur sur la moyenne quand σ est connu. Alors :

$$\mathbb{P}(\bar{x} - E < \mu < \bar{x} + E) = 2\phi(z_{\alpha/2}) - 1 = \begin{cases} 0,95 & \text{si } z_{\alpha/2} = 1,96 \\ 0,9 & \text{si } z_{\alpha/2} = 1,645 \end{cases}$$

Pour $z_{\alpha/2} = 1,96$ on pourra écrire que « μ appartient à l'intervalle de confiance $[\bar{x} - E; \bar{x} + E]$ au niveau $0,95$ ou bien au seuil de risque de 5% ».
On pourra aussi écrire : « on a confiance à 95% que cet intervalle contient la moyenne μ de la population ».

Comment déterminer un intervalle de confiance quand on ignore σ ?

- A partir de votre échantillon aléatoire simple, vous déterminez $\bar{x}$ et s.
- Vous vérifiez la normalité de votre échantillon (approximativement et dans ce cas il suffit d'avoir $n \geq 5$) ou bien vous avez plus de 30 individus.
- s servira d'estimateur pour σ mais cette approximation accroît l'incertitude... aussi il nous faut des intervalles de confiance plus grand pour pouvoir assurer que dans $(1-\alpha)\%$ des cas, la moyenne μ est dans l'intervalle de confiance. On remplace donc $z_{\alpha/2}$ par le coefficient $t_{\alpha/2}$ de la loi de Student à $n-1$ degrés de liberté[2] pour laquelle la densité est plus plate et dépend de la taille de l'échantillon.
- L'intervalle de confiance devient :
$$IC = [\bar{x} - E; \bar{x} + E]$$
où $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$ avec $t_{\alpha/2}$ a $n-1$ degrés de libertés

[2]table en annexe A.3

Exemple : En analysant un échantillon aléatoire simple de 40 grains issus d'une surface de sédimentation, nous cherchons cette fois à estimer le diamètre moyen des grains dans la population.

Imaginons qu'on obtienne $\bar{x} = 0,6mm$ et $s = 0,4mm$.
Alors, le nombre $n = 40$ de grains étant suffisant, on peut utiliser les coefficients de la loi de Student à $40 - 1 = 39$ degrés de liberté. A l'aide de la table fournie en annexe, on obtient pour $\alpha = 0.05$ la valeur $t_{\alpha/2} = 2,024$
La formule ci-dessus permet alors d'obtenir : $E = 0,128$
Conclusion : L'intervalle de confiance est $[0,6 - 0,128; 0,6 + 0,128] = [0,472; 0,728]$ qu'on arrondi à $[0,4; 0,7]$. A partir de notre échantillon, **on peut être sûr à** 95% **que l'intervalle de confiance contient la moyenne** μ **des diamètres de grains de cette surface sédimentaire**.