Statistiques inférentielles : estimation

Table des matières

Les problèmes de l'échantillonnage et de l'estimation sont illustrés par l'étude de la situation suivante :

Exemple 1
Un industriel produit en très grand nombres des yaourts, pour lesquelles l'usinage doit respecter des normes sanitaires draconiennes. Á la suite de mauvais réglages de l'une des machines, l'industriel a produit 1 million de ces yaourts, dont beaucoup risquent ainsi de présenter des dangers pour le consommateur.
Il souhaite connaître la proportion de yaourts susceptibles de rendre malade un client, afin de savoir s'il doit détruire sa production, ce qui représentera un fort manque à gagner, ou s'il peut malgré courir le risque de quelques gênes isolées dans la population, sans craindre de campagne médiatique mettant en cause ces yaourts, ce qui lui causerait un préjudice encore plus grand.
Il est ainsi prêt à détruire son stock ainsi produit si la proportion de yaourts dangereux pour la santé dépasse les $0,01\%$ de sa production.
Il n'est bien entendu pas question d'analyser un par un tous les yaourts produits : cela lui reviendrait encore plus cher, et de toutes façons, il faudrait ouvrir les yaourts, ce qui les rendrait invendables. Il décide donc d'effectuer un sondage c'est à dire de prélever par exemple 100 yaourts, de les faire analyser, et de relever la proportion de yaourts contaminés dans cet échantillon.
Il obtient ainsi la résultat suivant : dans l'échantillon prélevé (au hasard) parmi les yaourts produits, on en a trouvé 2% qui contenaient des germes. Notre industriel est-il plus avancé après ces analyses pour résoudre son problème ?

La réponse est bien sûr négative : en effet, il peut toujours se poser les questions suivantes :

1. aurait-on obtenu le même pourcentage en prélevant un autre échantillon ? (autrement dit, la proportion inquiétante relevée dans le premier échantillon est-elle due à de la malchance ?)
2. l'analyse de 100 yaourts sur le million produit est-elle suffisante ?
3. quelle confiance peut-on accorder au fait que l'analyse d'un échantillon de 100 yaourts ait conduit à une proportion de 2% de produits contaminés ?
4. aurait-on gagné en fiabilité du pronostique si l'on en avait fait analyser 200, 1000, 10000 yaourts ?

La question 1., elle, relève du champ de l'échantillonnage. Cette théorie répond à la question : "comment varie la proportion relevée d'un échantillon à l'autre, sachant que tous sont de même taille donnée à l'avance ?". Ces questions ont des réponses fournies par le théorème de la limite centrée vue dans un précédent chapitre. Les questions 2., 3. et 4., portant sur la taille de l'échantillon, et sur la confiance que l'on peut accorder au sondage sont du domaine de l'estimation : elles obtiennent une réponse avec les résultats sur la "loi des grands nombres".

I Estimation ponctuelle d'un paramètre

I.1 Moyenne

Propriété 1
La valeur moyenne m_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation $\overline{x}$ de la moyenne réelle de ce paramètre sur la population considérée.

Exemple 2
Une usine produit des vis cruciformes. On souhaite estimer la moyenne des longueurs des vis dans la production de la journée qui s'élève à 10000 pièces.
On choisit un échantillon de 150 vis et on obtient une moyenne de $m_e = 4,57$ cm.
On en déduit donc que la longueur moyenne des vis de la production journalière est $\overline{x} = 4,57$ cm.

I.2 Écart-type

Le problème est toujours le même, mais cette fois-ci, l'estimation de l'écart-type est moins intuitive ...

Propriété 2
L'écart-type σ_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation faussée de l'écart-type de ce paramètre dans toute la population considérée.
Une meilleure estimation σ de l'écart-type réel est obtenue en considérant le nombre $\sigma = \sigma_e\sqrt{\frac{n}{n-1}}$, où n est la taille de l'échantillon servant au calcul de σ_e.

Exemple 3
La mesure de la longueur des vis produites dans l'échantillon précédent de 150 pièces conduit à relever un écart-type de 3 mm.
La meilleure estimation possible de l'écart-type de la production journalière n'est pas de 3 mm comme dans le cas précédent pour la moyenne, mais de $\sigma = 3\sqrt{\frac{150}{149}} \simeq 3,01$ mm.

Remarque 1
La correction devient assez rapidement minime lorsque la taille de l'échantillon augmente car
$\lim\limits_{n\to\infty} \sqrt{\frac{n}{n-1}} = 1$.
La correction est ainsi de l'ordre de $0,5\%$ pour des échantillons de taille 100, et de l'ordre de $0,05\%$ pour des échantillons de taille 1000.

I.3 Fréquence

Propriété 3
La fréquence d'apparition f_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation f de la fréquence réelle d'apparition de ce paramètre sur la population considérée.

Exemple 4
Dans l'exemple précédent, On prélève un échantillon de 150 vis et on relève 3 pièces défectueuses.
On peut alors donner une estimation de la fréquence f de vis défectueuses dans la production journalière :
On a $f_e = \frac{3}{150} = 0,02$ donc, $f = 0,02$.

Remarque 2
Notons qu'il revient exactement au même d'estimer un pourcentage : dans l'exemple précédent, on peut affirmer que 2% des vis ont une croix mal formée sur la tête.

II Estimation par intervalle de confiance d'un paramètre

Les estimations ponctuelles proposées ci-dessus dépendent directement de l'échantillon prélevé au hasard. Dans de très nombreux cas, l'importance attribuée au hasard est grande, cela conduit à s'interroger avant d'utiliser ces estimations pour prendre des décisions dont les conséquences peuvent être lourdes !
Aussi, sans rejeter les informations fournies par l'étude d'un échantillon, est-on amené à chercher un nouveau type d'estimation de la fréquence et de la moyenne d'une population, en utilisant le calcul de probabilités qui permet de « contrôler » l'influence d'un échantillon particulier.

II.1 Moyenne

On souhaite, à partir des observations faites sur un échantillon, déterminer un intervalle de confiance contenant la valeur moyenne avec un risque d'erreur décidé à l'avance.
On suppose que les conditions sont réunies pour faire l'approximation que la loi d'échantillonnage de la moyenne $\overline{X}$ est la loi normale $\mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$.

On pose $T = \dfrac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}}$, T suit donc la loi normale centrée réduite $\mathcal{N}(0;1)$.

Soit α la probabilité, fixée à l'avance, pour que T n'appartienne pas à l'intervalle $[-t;t]$, on peut écrire :

$$\begin{aligned}
P(|T| > t) = \alpha &\iff 1 - P(|T| \leq t) = \alpha \\
&\iff P(|T| \leq t) = 1 - \alpha \\
&\iff P(-t \leq T \leq t) = 1 - \alpha \\
&\iff P\left(-t \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq t\right) = 1 - \alpha \\
&\iff P\left(\overline{X} - t\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + t\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha
\end{aligned}$$

Autrement dit, m appartient à l'intervalle $\left[\overline{X} - t\frac{\sigma}{\sqrt{n}}; \overline{X} + t\frac{\sigma}{\sqrt{n}}\right]$ pour $100(1-\alpha)\%$ des échantillons.

- Cet intervalle est appelé intervalle de confiance,
- α est le risque d'erreur ou le seuil de risque,
- $1 - \alpha$ est le coefficient de confiance.

Propriété 4
L'intervalle $\left[\overline{X} - t\frac{\sigma}{\sqrt{n}}; \overline{X} + t\frac{\sigma}{\sqrt{n}}\right]$ est l'intervalle de confiance de la moyenne m de la population avec le coefficient de confiance $2\Pi(t) - 1 = 1 - \alpha$.

$$2\Pi(t) - 1$$

Remarque 3
Les valeurs fréquentes du niveau de confiance sont $0,99$ et $0,95$.
Pour ces deux valeurs, on obtient successivement $t = 2,575$ et $t = 1,96$.

Exemple 5
On suppose que la durée de vie, exprimée en heures, d'une ampoule électrique d'un certain type, suit la loi normale de moyenne M inconnue et d'écart-type $\sigma = 20$.
Une étude sur un échantillon de 16 ampoules donne une moyenne de vie égale à 3000 heures.
On va déterminer un intervalle de confiance de M au seuil de risque de 10%.
On a : $\alpha = 10\%$ d'où $2\Pi(t) - 1 = 0,90 \iff \Pi(t) = 0,95 \iff t = 1,645$.
Un intervalle de confiance de M est donc : $\left[3000 - 1,645\frac{20}{\sqrt{16}}; 3000 + 1,645\frac{20}{\sqrt{16}}\right] = [2992, 3008]$.

II.2 Fréquence

A l'aide d'un échantillon, nous allons définir, avec un coefficient de confiance choisi à l'avance, un intervalle de confiance de la fréquence p des éléments de la population possédant une certaine propriété.

On se place dans le cas où on peut approximer la loi par la loi normale $\mathcal{N}\left(p; \sqrt{\dfrac{p(1-p}{n}}\right)$.

Propriété 5
L'intervalle $\left[f - t\sqrt{\dfrac{f(1-f)}{n-1}}; f + t\sqrt{\dfrac{f(1-f)}{n-1}}\right]$ est l'intervalle de confiance d'une fréquence p de la population avec le coefficient de confiance $2\Pi(t) - 1 = 1 - \alpha$ ayant pour centre la fréquence f de l'échantillon considéré.

Exemple 6
Un sondage dans une commune révèle que sur les 500 personnes interrogées, 42% sont mécontentes de l'organisation des transport. On veut déterminer, au seuil de risque 1%, un intervalle de confiance du pourcentage p de personnes mécontentes dans la commune :
On a : $f = 0,42$; $n = 500$; $\alpha = 1\%$ donc $t = 2,575$.
Un intervalle de confiance du pourcentage p est donc :
$\left[0,42 - 2,58\sqrt{\dfrac{0,42 \times 0,58}{499}}; 0,42 + 2,575\sqrt{\dfrac{0,42 \times 0,58}{499}}\right] = [0,36; 0,48] = [36\%; 47\%]$.

III Tableau récapitulatif

Le tableau ci-dessous regroupe toutes les situations dans lesquelles on doit savoir fournir une estimation ponctuelle ou par intervalle de confiance :

Paramètre de la population totale à estimer	Valeur du paramètre dans l'échantillon de taille n	Estimation ponctuelle pour la population totale	Estimation par intervalle de confiance au niveau de confiance $2\Pi(t) - 1$ pour la population totale
Moyenne	m_e	$m = m_e$	$\left[m_e - t\dfrac{\sigma}{\sqrt{n}}; m_e + t\dfrac{\sigma}{\sqrt{n}}\right]$
Écart-type	σ_e	$\sigma = \sigma_e\sqrt{\dfrac{n}{n-1}}$	
Fréquence	f_e	$f = f_e$	$\left[f_e - a\sqrt{\dfrac{f_e(1-f_e)}{n-1}}; f_e + t\sqrt{\dfrac{f_e(1-f_e)}{n-1}}\right]$