Échantillonnage et estimation

L'objectif du chapitre est de forger des outils de statistique, dont les applications sont maintenant largement diffusées dans les processus industriels de production les plus performants, en particulier en contrôle de qualité. Les applications des outils que nous allons mettre en place couvrent aussi des domaines plus larges, englobant la théorie des sondages, et la mise en place d'outils d'aide à la prise de décision, élaborés à partir d'outils statistiques sur divers indicateurs.
La plupart des résultats de ce chapitre ont constitué un des cœurs de l'activité de recherche mathématiques du XX° siècle, dans les domaines des Probabilités et des Statistiques : la majorité des outils exposés ont moins de 150 ans, ce qui est tout-à-fait remarquable, et montre, outre l'intérêt d'entreprendre des recherches scientifiques, la richesse des résultats auxquels elles parviennent.

1 Échantillonnage et estimation : principes généraux et objets d'étude

Les problèmes de l'échantillonnage et de l'estimation sont illustrés par l'étude de la situation suivante :
Activité. 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 ? 10.000 ?
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".
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. De l'activité précédente, retenons l'essentiel ; les théories de l'échantillonnage et de l'estimation, bien que liées, diffèrent dans leurs objets d'étude :
- en échantillonnage, on connaît au départ les caractéristiques d'une population ; on cherche alors à connaître les effets du prélèvement (aléatoire) d'un nombre donnée d'individus dans cette population sur cette proportion : dans un échantillon, cette proportion sera-t-elle respectée exactement ? ou avec quelle variation par rapport à la proportion initiale ?

- en estimation, on dispose d'informations parcellaires sur une population au moyen d'un sondage, et on on veut étendre ces résultats pour décrire toute la population : dans quelle mesure cela est-il possible ? avec quelle précision ? avec quelle confiance ?

Exemple 1 *On sait que dans la population française, il y a* 45% *d'hommes au dernier recensement. On interroge au hasard* 1000 *français lors de* 30 *enquêtes : combien d'enquêtes ont conduit à observer exactement ce pourcentage d'hommes ? combien d'enquêtes ont conduit à observer un pourcentage d'hommes entre* 40% *et* 50% *? Quel est en moyenne le pourcentage d'hommes relevés durant ces enquêtes ?*
Ces questions relèvent de la théorie de l'échantillonnage ; le théorème de la limite centrée fournit des réponses.

Exemple 2 *On a réalisé un sondage auprès de* 1000 *français, et obtenu une proportion de* 63% *de gens qui veulent changer de Premier Ministre : combien de Français souhaitent réellement changer de Premier Ministre ? quelle confiance accorder à ce sondage ? a-t-on pris l'avis d'un nombre suffisant de Français pour prendre une décision ?*
Ces questions relèvent de la théorie de l'estimation ; la loi des grands nombres fournit des réponses.

2 Échantillonnage

2.1 Fluctuation d'échantillonnage

Dans le premier paragraphe de ce cours, on a observé les résultats de lancers de pièces à Pile ou Face. On a constaté dans le chapitre précédent que sur des échantillons comparables (constitués de 20 lancers successifs, tout comme sur des échantillons de 200 lancers), que la composition en Pile et en Face de chacun de ces échantillons n'est pas homogène : d'un échantillon à l'autre, il n'y a pas toujours le même nombre de Pile et de Face ... car les échantillons sont obtenus après un tirage au sort : même si la pièce lancée est parfaitement équilibrée, il n'y a pas de raison de n'obtenir que des échantillons composés à parité de Pile et de Face.
La question qui de pose est alors la suivante : comment se comporte la répartition de Pile et de Face d'un échantillon à l'autre ? Autrement dit, dans quelle mesure des calculs de moyennes ou de fréquences empiriques, effectués à partir de divers échantillons conduisent-ils à des valeurs différentes ? Sont-ils "rigides" ? aléatoires ? imprévisibles ?
Ce paragraphe présente deux résultats très simples, qui répondent en partie à cette question : simplement parlant, lorsqu'on tire un grand nombre d'échantillons de même taille, on observe une répartition tout-à-fait régulière (et suivant une loi normale) des fréquences empiriques d'apparition du Pile ...

2.2 Comportement des moyennes empiriques calculées sur les échantillons

Théorème 1 *Soit une population dont une caractéristique X suit une loi de moyenne m et d'écart-type σ. Soit M_n la variable aléatoire qui, à un échantillon de taille n prélevé dans cette population, associe la moyenne empirique de cette caractéristique dans cet échantillon.*
Alors, si n est assez grand, M_n suit une loi normale $\mathcal{N}\left(m; \dfrac{\sigma}{\sqrt{n}}\right)$.

Preuve
◄ Ce théroème est une reformulation du théorème de la limite centrée. En effet, l'observation d'un échantillon de taille n consiste en l'observation de la répétition de n expériences aléatoires, qui peuvent être supposées indépendantes si n est grand (tirage avec remise). L'expérience i amène à observer la valeur X_i de X ($1 \leq i \leq n$), qui est une variable aléatoire de même loi que X.
Dans ces conditions, $M_n = \dfrac{X_1 + X_2 + ... + X_n}{n}$, et le théorème de la limite centrée indique que, si n est assez grand, M_n suit un loi normale $\mathcal{N}\left(m; \dfrac{\sigma}{\sqrt{n}}\right)$. ►

Remarque 1 *En pratique, pour $n \geq 30$, on peut appliquer le théorème précédent.*

Remarque 2 *Ce théorème indique en particulier :*

i. que les échantillons donnent des moyennes empiriques de la caractéristique X distribuées suivant une loi normale ;

ii. que les échantillons conduisant à une moyenne empirique de X proche de la moyenne de X dans la population étudiée sont les plus probables ;

iii. que les échantillons conduisant à une moyenne empirique de X proche de la moyenne de X dans la population étudiée sont d'autant plus souvent rencontrés que les échantillons sont de grande taille.

Exemple 3 *Au lycée, on a interrogé tous les élèves pour connaître leur taille. On a ainsi appris que la taille moyenne des élèves est de $1,70$ mètres, avec un écart-type de $0,1$. On s'apprête à constituer un groupe de n élèves, choisis au hasard, et à calculer leur taille moyenne. On cherche la probabilité que la moyenne obtenue à partir de l'observation de ce groupe se situe entre $1,65$ m et $1,75$m, pour $n = 10$, $n = 50$ et $n = 100$.*
Soit T la taille moyenne (en mètres) des n individus constituant l'échantillon prélevé. Le théorème précédent affirme que T est une variable aléatoire, qui suit une loi normale $\mathcal{N}\left(1,7; \dfrac{0,1}{\sqrt{n}}\right)$.

<u>Si $n = 10$</u> : T est une variable aléatoire qui suit une loi $\mathcal{N}\left(1,7; \dfrac{0,1}{\sqrt{10}}\right)$, soit approximativement une loi $\mathcal{N}(1,7; 0,0316)$.
Donc

$$\mathbb{P}(T \in [1,65; 175]) = \mathbb{P}(1,65 \leq T \leq 1,75) = \mathbb{P}\left(\frac{1,65-1,7}{0,0316} \leq \frac{T-1,7}{0,0316} \leq \frac{1,75-1,7}{0,0316}\right).$$

Posons $T^* = \dfrac{T-1,7}{0,0316}$; alors T^* suit par hypothèse une loi $\mathcal{N}(0;1)$, et on a donc

$$\mathbb{P}(T \in [1,65; 1,75]) \simeq \mathbb{P}(-1,58 \leq T^* \leq 1,58) = 2\Pi(1,58) - 1 = 2 \times 0,9429 - 1 \simeq 0,8858.$$

Il y a donc a priori $88,58\%$ des échantillons de taille 10 qui conduisent à l'observation d'une taille moyenne des élèves comprise entre $1,65$ m et $1,75$ m, dans cet échantillon.

<u>Si $n = 50$</u> : T est une variable aléatoire qui suit une loi $\mathcal{N}\left(1,7; \dfrac{0,1}{\sqrt{50}}\right)$, soit approximativement une loi $\mathcal{N}(1,7; 0,0141)$.
Donc

$$\mathbb{P}(T \in [1,65; 175]) = \mathbb{P}(1,65 \leq T \leq 1,75) = \mathbb{P}\left(\frac{1,65-1,7}{0,0141} \leq \frac{T-1,7}{0,0141} \leq \frac{1,75-1,7}{0,0141}\right).$$

Posons $T^* = \dfrac{T-1,7}{0,0141}$; alors T^* suit par hypothèse une loi $\mathcal{N}(0;1)$, et on a donc

$$\mathbb{P}(T \in [1,65; 1,75]) \simeq \mathbb{P}(-3,54 \leq T^* \leq 3,54) = 2\Pi(3,54) - 1 = 2 \times 0,9998 - 1 \simeq 0,9996.$$

Il y a donc a priori $99,96\%$ des échantillons de taille 50 qui conduisent à l'observation d'une taille moyenne des élèves comprise entre $1,65$ m et $1,75$ m, dans cet échantillon.

<u>Si $n = 100$</u> : T est une variable aléatoire qui suit une loi $\mathcal{N}\left(1,7; \dfrac{0,1}{\sqrt{100}}\right)$, soit approximativement une loi $\mathcal{N}(1,7; 0,01)$.
Donc

$$\mathbb{P}(T \in [1,65; 175]) = \mathbb{P}(1,65 \leq T \leq 1,75) = \mathbb{P}\left(\frac{1,65-1,7}{0,01} \leq \frac{T-1,7}{0,01} \leq \frac{1,75-1,7}{0,01}\right).$$

Posons $T^* = \dfrac{T-1,7}{0,0316}$; alors T^* suit par hypothèse une loi $\mathcal{N}(0;1)$, et on a donc

$$\mathbb{P}(T \in [1,65; 1,75]) \simeq \mathbb{P}(-5 \leq T^* \leq 5) = 2\Pi(5) - 1 = 2 \times 1 - 1 = 1$$

A priori, pratiquement tous les échantillons de taille 100 conduisent à l'observation d'une taille moyenne des élèves comprise entre $1,65$ m et $1,75$ m, dans cet échantillon.

2.3 Comportement des proportions empiriques calculées sur les échantillons

Donnons un corollaire utile du théorème précédent, dans le cas où l'on observe un caractère de fréquence donnée dans une population, et où l'on souhaite connaître la répartition des valeurs empiriquement calculées de cette fréquence :

Théorème 2 *Soit une population dont un caractère est présent en proportion p. Soit F_n la variable aléatoire qui, à un échantillon de taille n prélevé dans cette population associe la proportion empirique d'apparition du caractère dans cet échantillon.*
Alors, si n est assez grand, F_n suit une loi $\mathcal{N}\left(p; \sqrt{\dfrac{p(1-p)}{n}}\right)$.

Preuve
◄ Ce théroème est une reformulation du théorème de la limite centrée. En effet, si X désigne la variable alétoire qui vaut 1 si le caractère est observé lors du prélèvement d'un individu de la population, et 0 sinon. l'observation d'un échantillon : X suit une loi $\mathcal{B}(1;p)$, donc $\mathbb{E}(X) = p$ et $\sigma_X = \sqrt{p(1-p)}$.
Un échantillon de taille n consiste en l'observation de la répétition de n expériences aléatoires, qui peuvent être supposées indépendantes si n est grand (tirage avec remise). L'expérience i amène à observer la valeur X_i de X ($1 \leq i \leq n$), qui est une variable aléatoire de même loi que X.
Dans ces conditions, $F_n = \dfrac{X_1 + X_2 + ... + X_n}{n}$ est la fréquence d'apparition du caractère dans un échantillon de taille n, et le théorème de la limite centrée indique que, si n est assez grand, F_n suit une loi normale $\mathcal{N}\left(\mathbb{E}(X); \dfrac{\sigma_X}{\sqrt{n}}\right)$. Donc F_n suit une loi normale $\mathcal{N}\left(p; \sqrt{\dfrac{p(1-p)}{n}}\right)$. ►

Remarque 3 *En pratique, pour $n \geq 30$, on peut appliquer le théorème précédent.*

Remarque 4 *Ce théorème indique en particulier :*

i. *que les échantillons conduisent à des fréquences empiriques d'apparition du caractère distribuées suivant une loi normale ;*
ii. *que les échantillons qui ont la même composition (relativement au caractère étudié) que la population étudiée sont les plus probables ;*
iii. *que les échantillons qui ont la même composition (relativement au caractère étudié) que la population étudiée sont d'autant plus souvent rencontrés que les échantillons sont de grande taille.*

Exemple 4 *Afin d'illustrer ce théorème, on a analysé* 10.000 *échantillons de taille* 20 *(donc constitués de* 20 *lancers de pièce à Pile ou Face), et on a, pour chaque échantillon, relevé la fréquence empirique d'apparition du Pile.*
Puis on a dressé le tableau suivant, qui donne une idée de la répartition des 10.000 *fréquences empiriques calculées à partir des divers échantillons. Enfin, ce tableau présente un rappel de la densité (discrétisée) de la*

loi normale qui rend théoriquement compte de la distribution théorique de ces fréquences empiriques.

Intervalles de fréquences empiriques	*Proportion d'échantillons présentant cette fréquence empirique*	*Proportion théorique selon une loi* $\mathcal{N}\left(\frac{1}{2}; \frac{\sqrt{\frac{1}{2}(1-\frac{1}{2})}}{\sqrt{20}}\right)$
$[0; 0,05[$	0	0
$[0,05; 0,10[$	$0,0001$	$0,0001$
$[0,10; 0,15[$	$0,0011$	$0,0007$
$[0,15; 0,20[$	$0,0046$	$0,0028$
$[0,20; 0,25[$	$0,0147$	$0,0090$
$[0,25; 0,30[$	$0,0343$	$0,0241$
$[0,30; 0,35[$	$0,0714$	$0,0530$
$[0,35; 0,40[$	$0,1229$	$0,0957$
$[0,40; 0,45[$	$0,1637$	$0,1418$
$[0,45; 0,50[$	$0,1741$	$0,1726$
$[0,50; 0,55[$	$0,1652$	$0,1726$
$[0,55; 0,60[$	$0,1212$	$0,1418$
$[0,60; 0,65[$	$0,0721$	$0,0957$
$[0,65; 0,70[$	$0,0365$	$0,0530$
$[0,70; 0,75[$	$0,0139$	$0,0241$
$[0,75; 0,80[$	$0,0028$	$0,0090$
$[0,80; 0,85[$	$0,0010$	$0,0028$
$[0,85; 0,90[$	$0,0002$	$0,0007$
$[0,90; 0,95[$	$0,0002$	$0,0001$
$[0,95; 1[$	0	0

Exemple 5 *Une machine produit* 10% *de pièces défectueuses. On cherche à déterminer, pour* $n = 10$, $n = 100$, *et* $n = 1000$, *la probabilité qu'un échantillon de* n *pièces produites amène à constater, dans cet échantillon, une proportion de pièces défectueuses comprise entre* 8% *et* 12%.
Soit F *la proportion de pièces défectueuses dans un échantillon de taille* n. *Le théorème précédent affirme que* F *est une variable aléatoire, qui suit une loi normale* $\mathcal{N}\left(0,1; \sqrt{\dfrac{0,1 \times 0,9}{n}}\right)$.

Si $n = 10$ *:* F *est une variable aléatoire qui suit une loi* $\mathcal{N}\left(0,1; \sqrt{\dfrac{0,1 \times 0,9}{10}}\right)$, *soit approximativement une loi* $\mathcal{N}(0,1; 0,095)$.
Donc

$$\mathbb{P}(F \in [0,08; 0,12]) = \mathbb{P}(0,08 \leq F \leq 0,12) = \mathbb{P}\left(\frac{0,08 - 0,1}{0,095} \leq \frac{F - 0,1}{0,095} \leq \frac{0,12 - 0,1}{0,095}\right).$$

Posons $F^* = \dfrac{F - 0,1}{0,095}$ *; alors* F^* *suit par hypothèse une loi* $\mathcal{N}(0;1)$, *et on a donc*

$$\mathbb{P}(F \in [0,08; 0,12]) \simeq \mathbb{P}(-0,21 \leq F^* \leq 0,21) = 2\Pi(0,21) - 1 = 2 \times 0,5832 - 1 \simeq 0,1664.$$

Il y a donc a priori $16,54\%$ *des échantillons de taille* 10 *qui conduisent à l'observation d'une proportion de pièces défectueuses comprise entre* 8% *et* 12%.

Si $n = 100$ *:* F *est une variable aléatoire qui suit une loi* $\mathcal{N}\left(0,1; \sqrt{\dfrac{0,1 \times 0,9}{100}}\right)$, *soit une loi* $\mathcal{N}(0,1; 0,03)$.
Donc

$$\mathbb{P}(F \in [0,08; 0,12]) = \mathbb{P}(0,08 \leq F \leq 0,12) = \mathbb{P}\left(\frac{0,08 - 0,1}{0,03} \leq \frac{F - 0,1}{0,03} \leq \frac{0,12 - 0,1}{0,03}\right).$$

Posons $F^* = \dfrac{F - 0,1}{0,03}$ *; alors* F^* *suit par hypothèse une loi* $\mathcal{N}(0;1)$, *et on a donc*

$$\mathbb{P}(F \in [0,08; 0,12]) \simeq \mathbb{P}(-0,67 \leq F^* \leq 0,67) = 2\Pi(0,67) - 1 = 2 \times 0,7486 - 1 \simeq 0,4972.$$

Il y a donc a priori $49,72\%$ *des échantillons de taille* 100 *qui conduisent à l'observation d'une proportion de pièces défectueuses comprise entre* 8% *et* 12%.

Si $n = 1000$: F est une variable aléatoire qui suit une loi $\mathcal{N}\left(0,1;\sqrt{\frac{0,1\times 0,9}{1000}}\right)$, soit approximativement une loi $\mathcal{N}(0,1;0,0095)$.
Donc

$$\mathbb{P}(F\in[0,08;0,12]) = \mathbb{P}(0,08\leq F\leq 0,12) = \mathbb{P}\left(\frac{0,08-0,1}{0,0095}\leq\frac{F-0,1}{0,0095}\leq\frac{0,12-0,1}{0,095}\right).$$

Posons $F^* = \frac{F-0,1}{0,0095}$; alors F^* suit par hypothèse une loi $\mathcal{N}(0;1)$, et on a donc

$$\mathbb{P}(F\in[0,08;0,12]) \simeq \mathbb{P}(-2,1\leq F^*\leq 2,1) = 2\Pi(2,1)-1 = 2\times 0,9821-1 \simeq 0,9642.$$

Il y a donc a priori $96,42\%$ des échantillons de taille 1000 qui conduisent à l'observation d'une proportion de pièces défectueuses comprise entre 8% et 12%.

3 Estimation

Dans ce paragraphe, la situation est toujours la même : plutôt que de faire le calcul des paramètres statistiques (moyenne, variance et écart-type) exacts d'une population, on calcule une estimation de ces paramètres sur un échantillon réduit de cette population. Ce paragraphe présente les résultats qui permettent de dire dans quelle mesure les valeurs calculées à partir de l'échantillon peuvent donner un aperçu valable des valeurs que l'on obtiendrait en menant une étude statistique exhaustive de la population.
Nous distinguons deux types d'estimations, de nature absolument différentes : les estimations ponctuelles, et les estimations par intervalle de confiance.

3.1 Estimations ponctuelles

3.1.1 Estimation ponctuelle d'une moyenne

Théorème 3 *La valeur moyenne m_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation $\widehat{m}$ valable de la moyenne réelle m de ce paramètre sur la population considérée.*

Exemple 6 *Une usine produit des vis cruciformes en grand nombre. On souhaite estimer la longueur moyenne des vis produites au cours de la journée. Pour cela, on sélectionne un échantillon de vis suffisamment important. La longueur moyenne des vis calculée sur cet échantillon fournit alors une estimation valable de la longueur des vis produites au cours de la journée.*
Plus précisément, parmi tous les calculs que l'on peut faire à partir de l'échantillon, celui qui consiste à calculer la longueur moyenne et à affirmer que ce résultat traduit la longueur moyenne des vis produites dans la journée, est le meilleur possible.
Il n'en reste pas moins vrai que le résultat que l'on attribue à la longueur moyenne de toutes les vis est probablement erroné : simplement, ce choix est en quelque sorte le "moins mauvais" [1].

3.1.2 Estimation ponctuelle d'une fréquence

Théorème 4 *La fréquence d'apparition f_e d'un paramètre observé sur un échantillon de population, dont la taille est fixée, fournit une estimation $\widehat{f}$ valable de la fréquence réelle f d'apparition de ce paramètre sur la population considérée.*

Exemple 7 *Une usine produit des vis cruciformes en grand nombre. On souhaite estimer la fréquence d'apparition du défaut D : "les vis produites ont une croix mal formée sur la tête" dans la production de la journée. Pour cela, on sélectionne un échantillon de vis suffisamment important. La fréquence d'apparition du défaut calculée sur cet échantillon fournit alors une estimation valable de la fréquence d'apparition du défaut D dans la production de la journée.*
Plus précisément, parmi tous les calculs que l'on peut faire à partir de l'échantillon, celui qui consiste à calculer la fréquence à parti de l'échantillon et à affirmer que ce résultat traduit la fréquence dans toute la production, est le meilleur possible.
Il n'en reste pas moins vrai que le résultat que l'on attribue à la fréquence de défaut au cours de la journée est

[1] on appelle estimateur tout algorithme de calcul conduisant à l'obtention d'une estimation. Le problème du statisticien est de déterminer, parmi tous les estimateurs possibles, celui, ou ceux qui présentent les meilleures capacités d'estimation fiable. On parle ainsi d'estimateur sans biais pour désigner un estimateur dont l'espérance d'erreur par rapport à la valeur exacte à estimer (calculée sur tous les échantillons possibles) est nulle. Ce n'est pas la seule qualité recherchée pour un estimateur ...

probablement erroné : simplement, ce choix est en quelque sorte le "moins mauvais" qui produise une valeur estimative unique.

Remarque 5 *Notons qu'il revient exactement au même d'estimer un pourcentage, comme le montre l'exemple suivant.*

Exemple 8 *Un contrôleur des impôts, fraîchement arrivé en Corse, souhaite connaître le pourcentage d'habitants ayant fraudé le fisc dans leur déclaration de revenu. En effet, il n'est pas fou, et ne veut pas se faire plastiquer inutilement sa voiture, et préfère, s'il devait y avoir peu de fraudeurs, passer une année tranquille à nager dans les calanques, plutôt que de risquer des représailles pour un résultat médiocre. Afin de savoir s'il trouvera beaucoup de fraudeurs ou pas, il contrôle systématiquement tous les habitants du premier village corse rencontré. Il consiste ainsi un échantillon. Ayant remarqué que* 80% *des habitants de ce village ont fraudé, pourra-t-il passer une année tranquille, ou devra-t-il accomplir sa mission ?*
Pour répondre, et en l'absence d'autres éléments objectifs, il faut bien faire confiance à la proportion relevée dans ce village. On est donc amené à considérer que dans tous les villages, il y aura environ 80% *de fraudeurs (et donc une année de travail intense, et en prime une voiture plastiquée ...). Les nationalistes diront,eux, qu'il s'agit d'une vision faussée de la Corse : "Cet inspecteur n'a pas eu de chance, et il est tombé sur le village qui rassemble tous les fraudeurs !"*
Ceci dit, l'inspecteur ne peut pas le savoir au départ ... et il fait logiquement le choix stratégique correspondant à la situation la plus vraissemblable à partir de sa maigre expérience.

3.1.3 Estimation ponctuelle d'un écart-type

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

Théorème 5 *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 $\widehat{\sigma}$ de l'écart-type σ réel est obtenue en considérant le nombre $\widehat{\sigma} = \sigma_e\sqrt{\dfrac{n}{n-1}}$, où n est la taille de l'échantillon servant au calcul de σ_e.

Exemple 9 *La mesure de la longueur des vis produites une heure consuit à relever un écart-type de* 2 *mm dans un échantillon prélevé constitué de* 40 *pièces. Alors la meilleure estimation possible de l'écart-type de la production journalière n'est pas de* 2 *mm, mais de* $2\sqrt{\frac{40}{39}} \simeq 2,03$ *mm.*

Remarque 6 *Notons que* $\sqrt{\dfrac{n}{n-1}} \to 1$ *lorque* $n \to +\infty$. *La correction devient assez rapidement minime lorsque la taille de l'échantillon augmente.*
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.

3.2 Estimations par intervalles de confiance

Comme on peut cependant s'en douter, la connaissance d'une estimation d'une moyenne ou d'une fréquence au moyen d'un seul nombre n'est pas satisfaisant, et ne constitue pas la panacée. Ce paragraphe explore une notion d'estimation plus précise, en ce sens que l'on se fixe à priori un niveau de confiance que l'on pourra accorder à l'estimation une fois qu'elle sera faite ; on détermine alors, non une seule valeur pour la moyenne ou la fréquence à estimer, mais un intervalle de valeurs qui contiendra, au niveau de confiance fixé, la valeur réelle de la grandeur à estimer. Il s'agit d'une estimation par intervalle de confiance.

3.2.1 Estimation par intervalles de confiance d'une moyenne : cas d'un écart-type connu pour le caractère à estimer

Théorème 6 *On cherche à connaître la valeur moyenne m d'un caractère quantitatif d'une population. On suppose que l'on connaît la valeur exacte de l'écart-type σ de ce caractère sur l'ensemble de la population.*
On considère un échantillon de taille n de cette population, et on suppose que, sur cet échantillon, la valeur moyenne empiriquement calculée du caractère est m_e.
Soit $a \in \mathbb{R}$ tel que $2\Pi(a) - 1$ soit le niveau de confiance recherché pour réaliser l'estimation.
Alors, avec une probabilité de $2\Pi(a)-1$, l'intervalle $\left[m_e - a\dfrac{\sigma}{\sqrt{n}}; m_e + a\dfrac{\sigma}{\sqrt{n}}\right]$ contient la valeur moyenne exacte du caractère, pour l'ensemble de la population considérée.

Exemple 10

Exemple 11

Remarque 7 *Un niveau de confiance de* $0,95$ *(soit* 95%*) ou de* $0,99$ *(soit* 99%*) signifie que la valeur exacte de la moyenne à estimer est située dans l'intervalle de confiance avec une probabilité de* $0,95$ *(ou* $0,99$*). Rien ne garantit donc absolument que l'intervalle de confiance contienne réellement la valeur à estimer ; tout au plus peut-on dire que le risque qu'il en soit autrement est faible ...*

Donnons maintenant une idée de la manière dont on peut démontrer ce théorème.
Preuve
◄ D'après les résultats sur la fluctuation d'échantillonnage, on sait que les moyennes empiriques calculées sur un échantillon de taille n prélevé dans une population soumise aux hypothèses du théorème suivent une loi $\mathcal{N}\left(m;\frac{\sigma}{\sqrt{n}}\right)$. Choisissons $a \in \mathbb{R}$ de sorte que $2\Pi(a)-1$ soit le niveau de confiance recherché. On a donc, puisque m_e est une réalisation de la variable aléatoire qui rend compte de la fluctuation des moyennes empiriques,

$$\begin{aligned}\mathbb{P}\left(m \in \left[m_e - a\frac{\sigma}{\sqrt{n}}; m_e + a\frac{\sigma}{\sqrt{n}}\right]\right) &= \mathbb{P}\left(m_e - a\frac{\sigma}{\sqrt{n}} \leq m \leq m_e + a\frac{\sigma}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(m - a\frac{\sigma}{\sqrt{n}} \leq m_e \leq m + a\frac{\sigma}{\sqrt{n}}\right)\\ &= \mathbb{P}\left(-a \leq \frac{(m_e - m)\sqrt{n}}{\sigma} \leq a\right)\\ &= 2\Pi(a) - 1,\end{aligned}$$

car alors $\dfrac{(m_e - m)\sqrt{n}}{\sigma} = \dfrac{m_e - m}{\frac{\sigma}{\sqrt{n}}}$ suit une loi $\mathcal{N}(0;1)$. On a donc montré que

$$\mathbb{P}\left(m \in \left[m_e - a\frac{\sigma}{\sqrt{n}}; m_e + a\frac{\sigma}{\sqrt{n}}\right]\right) = 2\Pi(a) - 1.$$

La moyenne m que l'on cherche à estimer est donc comprise dans l'intervalle de confiance défini plus haut avec une probabilité de $2\Pi(a) - 1$, comme annoncé. ►

3.2.2 Estimation par intervalles de confiance d'une moyenne : cas d'un écart-type inconnu pour le caractère à estimer

Théorème 7 *On cherche à connaître la valeur moyenne d'un caractère quantitatif sur l'ensemble d'une population.*
On considère un échantillon de taille n *de cette population, et on suppose que, sur cet échantillon, la valeur moyenne empiriquement calculée du caractère est* m_e*, avec un écart-type empiriquement calculé de* σ_e*.*
Soit $a \in \mathbb{R}$ *tel que* $2\Pi(a) - 1$ *soit le niveau de confiance recherché pour réaliser l'estimation.*
Alors, avec une probabilité de $2\Pi(a)-1$*, l'intervalle* $\left[m_e - a\frac{\sigma_e}{\sqrt{n-1}}; m_e + a\frac{\sigma_e}{\sqrt{n-1}}\right]$ *contient la valeur moyenne exacte du caractère, pour l'ensemble de la population considérée.*

Preuve
◄ Dans l'intervalle de confiance du théorème précédent, on a simplement remplacé l'écart-type σ du caractère sur l'ensemble de la population, dont la valeur est inconnue, par l'estimation ponctuelle de celui-ci, qui est $\sigma_e\sqrt{\dfrac{n}{n-1}}$. Cette substitution produit l'intervalle en question. ►

Remarque 8 *Il est à noter que, pour des échantillons assez important, la variation de l'amplitude de l'intervalle de confiance est négligeable : ceci se justifie par le fait que* $\dfrac{1}{\sqrt{n}} \simeq \dfrac{1}{\sqrt{n-1}}$ *pour* n *assez grand, et la loi des grands nombre, qui donne* $\sigma_e \simeq \sigma$ *pour* n *assez grand.*

Exemple 12

Exemple 13

3.2.3 Estimation par intervalles de confiance d'une fréquence

Théorème 8 *On cherche à connaître la fréquence d'apparition d'un caractère qualitatif sur l'ensemble d'une population.*
On considère un échantillon de taille n de cette population, et on suppose que, sur cet échantillon, la fréquence empiriquement calculée du caractère est f_e.
Soit $a \in \mathbb{R}$ tel que $2\Pi(a) - 1$ soit le niveau de confiance recherché pour réaliser l'estimation.
Alors, avec une probabilité de $2\Pi(a) - 1$, l'intervalle $\left[f_e - a\sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + a\sqrt{\frac{f_e(1-f_e)}{n-1}}\right]$ contient la fréquence exacte d'apparition du caractère, pour l'ensemble de la population considérée.

Exemple 14

Exemple 15

Donnons maintenant une idée de la manière dont on peut démontrer ce théorème.
Preuve
◀ D'après les résultats sur la fluctuation d'échantillonnage, on sait que les fréquences empiriques calculées sur un échantillon de taille n prélevé dans une population dans laquelle la fréquence exacte correspondante est f, suivent une loi $\mathcal{N}\left(f; \sqrt{\frac{f(1-f)}{n}}\right)$. Choisissons $a \in \mathbb{R}$ de sorte que $2\Pi(a) - 1$ soit le niveau de confiance recherché. On a donc, puisque f_e est une réalisation de la variable aléatoire qui rend compte de la fluctuation des fréquences empiriques,

$$\begin{aligned}
\mathbb{P}\left(f \in \left[f_e - a\sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + a\sqrt{\frac{f_e(1-f_e)}{n-1}}\right]\right) &= \mathbb{P}\left(f_e - a\sqrt{\frac{f_e(1-f_e)}{n-1}} \leq f \leq f_e + a\sqrt{\frac{f_e(1-f_e)}{n-1}}\right) \\
&= \mathbb{P}\left(0 \leq \left(\frac{f - f_e}{a}\right)^2 \leq \frac{f_e(1-f_e)}{n-1}\right) \\
&= \mathbb{P}\left(0 \leq \left(\frac{f_e - f}{a}\right)^2 \leq \frac{f_e(1-f_e)}{n-1}\right).
\end{aligned}$$

Or, dans terme inconnu $\frac{f_e(1-f_e)}{n}$, on peut considérer que f_e est une estimation ponctuelle de la fréquence f ; dans ce cas, $f_e = \frac{n-1}{n}f$, et donc $\frac{f_e(1-f_e)}{n-1} = \frac{f(1-\frac{n-1}{n})f}{n} \geq \frac{f(1-f)}{n}$. Ainsi, puisque $\frac{f(1-f)}{n} \leq \frac{f_e(1-f_e)}{n-1}$, on a $\left[0 \leq \left(\frac{f_e - f}{a}\right)^2 \leq \frac{f(1-f)}{n}\right] \subset \left[0 \leq \left(\frac{f_e - f}{a}\right)^2 \leq \frac{f_e(1-f_e)}{n-1}\right]$, ce qui entraîne la majoration

$$\mathbb{P}\left(0 \leq \left(\frac{f_e - f}{a}\right)^2 \leq \frac{f(1-f)}{n}\right) \leq \mathbb{P}\left(0 \leq \left(\frac{f_e - f}{a}\right)^2 \leq \frac{f_e(1-f_e)}{n-1}\right).$$

Par suite,

$$\begin{aligned}
\mathbb{P}\left(f \in \left[f_e - a\sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + a\sqrt{\frac{f_e(1-f_e)}{n-1}}\right]\right) &\geq \mathbb{P}\left(0 \leq \left(\frac{f_e - f}{a}\right)^2 \leq \frac{f(1-f)}{n}\right) \\
&= \mathbb{P}\left(f - a\sqrt{\frac{f(1-f)}{n}} \leq f_e \leq f + a\sqrt{\frac{f(1-f)}{n}}\right) \\
&= \mathbb{P}\left(f_e \in \left[f - a\sqrt{\frac{f(1-f)}{n}}; f + a\sqrt{\frac{f(1-f)}{n}}\right]\right) \\
&= 2\Pi(a) - 1.
\end{aligned}$$

La dernière égalité provient du fait que la fréquence f_e est distribuée suivant une loi $\mathcal{N}\left(f; \sqrt{\frac{f(1-f)}{n}}\right)$, d'après les résultats du paragraphe concernant l'échantillonnage. On a ainsi montré que

$$\mathbb{P}\left(f \in \left[f_e - a\sqrt{\frac{f_e(1-f_e)}{\sqrt{n-1}}}; f_e + a\sqrt{\frac{f_e(1-f_e)}{\sqrt{n-1}}}\right]\right) \geq 2\Pi(a) - 1.$$

La fréquence f que l'on cherche à estimer est donc comprise dans l'intervalle de confiance défini plus haut avec une probabilité de (au moins) $2\Pi(a) - 1$, comme annoncé. ►

3.3 Table des estimateurs usuels

Le tableau ci-dessous regroupe toutes les situations dans lesquelles on doit savoir fournir une estimation ponctuelle ou par intervalle de confiance à niveau donné d'un paramètre statistique d'une population, à partir d'un échantillon prélevé dans celle-ci.

Paramètre de la population totale à estimer	Valeur empirique du paramètre dans l'échantillon de taille n	Estimation ponctuelle pour la population totale	Estimation par intervalle au niveau de confiance $2\Pi(a) - 1$ pour la population totale
Moyenne	m_e	$\widehat{m} = m_e$	Écart-type σ de la population connu : $\left[m_e - a\frac{\sigma}{\sqrt{n}}; m_e + a\frac{\sigma}{\sqrt{n}}\right]$ Écart-type σde la population inconnu et écart-type σ_e de l'échantillon connu : $\left[m_e - a\frac{\sigma_e}{\sqrt{n-1}}; m_e + a\frac{\sigma_e}{\sqrt{n-1}}\right]$
Fréquence	f_e	$\widehat{f} = f_e$	$\left[f_e - a\sqrt{\frac{f_e(1-f_e)}{n-1}}; f_e + a\sqrt{\frac{f_e(1-f_e)}{n-1}}\right]$
Variance	$\mathbb{V}_e$	$\widehat{\mathbb{V}} = \frac{n}{n-1}\mathbb{V}_e$	
Écart-type	σ_e	$\widehat{\sigma} = \sigma_e\sqrt{\frac{n}{n-1}}$	

Notons pour compléter ce résumé que, pour un niveau de confiance de $0,95$, on choisit $a \simeq 1,96$, tandis que pour un niveau de confiance de $0,99$, on choisit $a \simeq 2,58$.

Table des matières