S3 - Échantillonnage

Maximilien Dreveton

May 6, 2017

1 Échantillonnage

1.1 Introduction

On étudie un caractère dans une population. La connaissance de la population entière n'est en général ni envisageable ni possible. Par contre, on peut connaître un échantillon de cette population (penser aux sondages : avec moins de 1 000 personnes, on en déduit les résultats d'une élection, et malgré les critiques, cela s'avère souvent être correct). Par contre, pour des produits de beauté, on a souvent des études "80% d'utilisateurs satisfait", suivi d'une petite astérisque précisant que l'étude a été effectuée sur 50 personnes. Intuitivement, on a la forte impression que 50 personnes ne sont pas suffisantes pour conclure.

L'objectif ici est donc d'estimer l'erreur faite lorsque l'on interroge uniquement n personnes.

Définition 1. *On appelle* ***échantillon de taille n*** *une liste de n résultats obtenus par n répétition indépendantes d'une même expérience aléatoire.*

Exemple 1 : On demande à n personnes pour qui ils vont voter; on regarde la durée de vie de n ampoules; résultat de n lancers de dé; etc...

Évidemment, si on prend deux échantillons de taille n, sauf cas particulier, les échantillons ne sont pas identiques: il y a des **fluctuations d'échantillonnage**.

Exemple : On lance un dé 1000 fois. Cela constitue un échantillon de taille 100. Si l'on relance 1000 fois le dé, on n'aura à priori pas exactement la même fréquence d'apparition de 1, même si cela ne devrait pas varier de beaucoup (à condition de définir ce que *beaucoup* signifie).

Dans toute la suite, on utilisera les notations suivantes :
- **p** = proportion du caractère dans l'ensemble de la population
- **f** = fréquence du caractère dans l'échantillon
- **n** = taille de l'échantillon

1.2 Théorème Central Limite

Théorème 1.1. *On étudie un caractère ayant une probabilité p d'être réalisé. On observe sa fréquence f d'apparition dans un échantillon de taille n.*

La probabilité que la fréquence du caractère observé appartienne à l'intervalle $I = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ *est de 95%.*

Proof. Largement hors programme, sera éventuellement démontré en terminale dans des cas simples. Le résultat écrit tel quel est d'ailleurs non optimal.

En effet, un meilleur intervalle est de $I = \left[p - \frac{1{,}96\sqrt{p(1-p)}}{\sqrt{n}}; p + \frac{1{,}96\sqrt{p(1-p)}}{\sqrt{n}}\right]$. On s'en sort car $0 \leq p \leq 1$, donc $p(1-p) \leq \frac{1}{4}$ donc $1{,}96\sqrt{p(1-p)} \leq \frac{1{,}96}{2} \leq 1$. Néanmoins, vous verrez plus tard une méthode plus élégante pour limiter la taille de cet intervalle de confiance.

De plus, le résultat est vrai dans la limite n grand. Pour éviter de se compliquer inutilement la vie, on prendra $n \leq 25$. Dans ce cas, il faut éviter les p trop proche de 0 ou de 1. On utilisera donc le théorème lorsque $0 \leq p \leq 0,8$, et $n \geq 25$.

Ce théorème est un corollaire du théorème central limite. □

Définition 2. *L'intervalle* $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ *est appelé intervalle de fluctuation de la fréquence f au seuil 95%.*

Cela veut dire que f appartient à cet intervalle avec une probabilité 95%

1.3 Première application : prise de décision sur un échantillon

Exemple 2 :

Dans un lac, 60% des espèces de poissons sont de la famille des saumons. On pèche au hasard 100 poissons; sur ces 100 poissons, on observe une fréquence f de saumons.

On a donc $p = 0,6$ et $n = 100$.

Un intervalle de fluctuation au seuil 0,95 de f est $I = [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}] = [0,6 - \frac{1}{\sqrt{1000}}; 0,6 + \frac{1}{\sqrt{1000}}] = [0,5; 0,7]$. Autrement dit, f a une probabilité 95% d'appartenir à l'intervalle $I = [0,5; 0,7]$.

Cela veut dire qu'en péchant 100 poissons, il y a 95% de chance que l'on pèche entre 50 et 70 saumons.

Exemple 3 : Une marque diffuse une publicité où elle mentionne que *70% de nos produits sont fabriqués en France*. La direction des fraudes prélève 1000 produits au hasard de la marque. Seulement 620 sont fabriqués en France. La marque a-t-elle menti dans son slogan publicitaire ?

Proof. Ici $p = 0,7$, $n = 1000$ et $f = \frac{620}{1000} = 0,62$. Un intervalle de confiance au seuil 0,95 pour f est $I = [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}] = [0,7 - \frac{1}{\sqrt{1000}}; 0,7 + \frac{1}{\sqrt{1000}}] = [0,67; 0,73]$.

Si le slogan de l'enseigne est juste, alors f appartient à I = [0,67 ; 0,73] avec une probabilité de 95%.

Or ici, $f = 0,62 \notin I$, et cela n'a que 5% de chance de se produire. Il y a donc forte probabilité pour que le slogan soit mensonger. □

Exemple 4 : Sur 100 lancer d'une pièce de monnaie, on obtient 41 piles. On nous affirme que la pièce n'est pas truquée. Que penser de cette affirmation ?

Proof. On a $p = 0,5$ (c'est l'affirmation *la pièce est non truquée*), $n = 100$ et $f = 0,41$.

L'intervalle de fluctuation au seuil 0,95 est $I = [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}] = [0,5 - \frac{1}{\sqrt{100}}; 0,5 + \frac{1}{\sqrt{100}}] = [0,4; 0,5]$

f appartient donc à l'intervalle de fluctuation I, donc l'affirmation *la pièce est non truquée* ne peut pas être rejetée. □

1.4 Deuxième application : déterminer p - Intervalle de fluctuation

En l'état, le théorème n'est pas utile. En effet, il est rarissime de connaître la probabilité p d'un caractère (c'est très souvent ce que l'on cherche justement !). Par contre, on connaît la fréquence f. On va le faire sur un exemple.

Exemple 5 : Élections présidentielle. Le jour du scrutin, on connaît la vraie probabilité p qu'un électeur ait voté pour un candidat A (cette probabilité est donc le vrai score du candidat A le jour de l'élection, modulo les abstentions et votes nuls et blancs). Par contre, avant le scrutin, on cherche à estimer cette probabilité, en demandant à 1 000 personnes pour qui ils vont voter : on récupère la fréquence f du nombre de personnes qui voteront pour le candidat A. Cette fréquence f est proche de p : c'est le principe des sondages. Mais il y a une marge d'erreur.

Question : quelle est cette marge d'erreur ?

Proof. On sait par le théorème 1.1 qu'avec une probabilité 95%, on a $f \in [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$

Or d'une part ;

$$f \leq p + \frac{1}{\sqrt{n}} \iff f - \frac{1}{\sqrt{n}} \leq p \tag{1}$$

et d'autre part :

$$p - \frac{1}{\sqrt{n}} \leq f \iff p \leq f + \frac{1}{\sqrt{n}} \tag{2}$$

Donc avec une probabilité 95%, on a

$$f - \frac{1}{\sqrt{n}} \leq p \leq f + \frac{1}{\sqrt{n}} \tag{3}$$

Pour revenir à l'exemple des sondages, on conclut que la marge d'erreur est de $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1000}} \approx 0,03 = 3\%$.

Autrement dit, la probabilité p que les gens votent pour le candidat A est comprise entre $f - 0,03$ et $f + 0,03$ avec une probabilité de 95%. (f étant le résultat du sondage).

□

Théorème 1.2. *On étudie un caractère ayant une probabilité p d'être réalisé. On observe sa fréquence f d'apparition dans un échantillon de taille n.*

La probabilité p de réalisation du caractère appartient à l'intervalle $I = [f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$ *avec une probabilité 0,95.*

Exemple 6 : Je lance une pièce de monnaie 1 000 fois, et j'obtiens pile 450 fois. Ainsi, la fréquence d'obtention de pile est de $f = \frac{450}{1000} \approx 0,45$. C'est proche de 0,5, donc à priori si la pièce est truquée, elle ne l'est pas de beaucoup.

Mais on peut dire plus avec le théorème précédent. En effet, l'intervalle de fluctuation de la fréquence f au seuil 95% est $I = [p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$

Exercice 1 : Le PDG d'une entreprise a pour objectif que 70% des clients soient satisfaits. Pour vérifier si c'est bien le cas, il organise une enquête auprès d'un échantillon de clients. Il interroge 200 clients. 148 se déclarent satisfaits.

Au vu de cet échantillon, l'objectif est-il atteint ?

Proof. La fréquence observée de satisfaction est $f = \frac{148}{200} = 0,74$.

Ici, p=0,75 et n=200. Un intervalle de fluctuation au seuil de 95% est :

$$I = [0,74 - \frac{1}{\sqrt{200}}; 0,74 + \frac{1}{\sqrt{200}}]$$

donc $I = [0.67; 0,81]$. Donc il y a 95% de chance que la proportion de clients satisfaits soit compris entre 67% et 81%. □

Exercice 2 : Un sondage pour l'élection présidentielle est réalisé sur 1 000 individus interrogés. Un candidat est crédité de 25% des intentions de vote. Il réalise finalement 29% des voix. Peut-on dire que le sondage s'est trompé ?

Proof.

$$I = [0,25 - \frac{1}{\sqrt{1000}}; 0,25 + \frac{1}{\sqrt{1000}}]$$

Donc $I = [0,22; 0,28]$. D'après les sondages, il y avait 95% de chance que le candidat fasse un score compris entre 22% et 28%. Pourtant, il réalise 29%. On peut donc considérer que les résultats du sondage étaient faux (ou du moins en dehors de la marge d'erreur, ce qui se produit avec une probabilité inférieure à 5%). □

Remarque : On vient de montrer qu'un sondage réalisé à partir de 1 000 individus a une marge d'erreur de 3%.

Proof. □

Exercice 3 : Une émission télévision a attiré 6 millions de téléspectateurs, soit 30% de part d'audience. Pour estimer le degré de satisfaction du public, une enquête est réalisée auprès de 300 personnes. Sur les 300 personnes, 105 ont regardé la série.

Ce sondage remet-il en question la part d'audience de 30% de cette série ?

1.5 Conclusion - Résumé

Les théorèmes 1.1 et 1.2 sont donc équivalents, mais ils ne s'appliquent pas dans les même conditions.

1. Théorème 1.1 : On connaît p et n; on a déterminer un échantillon, d'où l'on sort une fréquence f. A partir de cette fréquence, on peut valider l'échantillon.
   (a) Si la fréquence observée du caractère est dans l'intervalle de fluctuation, on *valide* l'échantillon.
   (b) Si la fréquence observées du caractère n'est pas dans l'intervalle de fluctuation, on **rejette** l'échantillon.

En pratique, rejeter l'échantillon revient souvent à dire que la probabilité donnée p est fausse (se référer à les exemples 1.3 et et 1.3). Mais cela pourrait aussi vouloir dire que l'échantillon est truqué.
Néanmoins, dans 5% des cas, la décision prise risque d'être incorrecte. C'est un risque faible que l'on accepte de prendre en toute conscience.

2. Théorème 1.2 : on connaît f et n via un sondage. Si n est très grand, on sait que $f \approx p$ (Loi des Grands Nombres vue dans le cours de proba S1). On veut donc avoir un intervalle de confiance pour p. On sait alors que p appartient à $I = [f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$ avec une probabilité de 95%.
Un physicien noterait $p = f \pm \frac{1}{\sqrt{n}}$.

1.6 (*) Compléments : Un ou des intervalles de confiance ?

Une étude menée fin 2008 sur 298 logements parisiens choisis au hasard dans l'annuaire assure que le prix du loyer au mètre carré est de 18.4 euros, avec un écart-type mesuré de 3.2 euros.

- L'observatoire des loyers parisiens annonce que pour 95% des logements, le prix des loyers au m2 est de $18,4 \pm 0,4$ euros (donc compris entre 18,0 et 18,8 euros).
- Pourtant, le collectif Jeudi Noir (association dénonçant les loyers trop cher), dénonce que 95% des loyers au m2 coûte plus de 18,1 euros.
- Enfin, l'association des propriétaires-bailleurs (qui au contraire aimerait que les loyers soient plus important), assure que les loyers sont tout à fait raisonnables, car 95% des loyers au m2 coûtent moins de 18,7 euros.

Explication

- L'observatoire des loyers parisiens utilise l'intervalle de confiance centrée autour de la moyenne 18,4 (obtenue après un sondage). Cet intervalle est $I_1 = [18,4 - \frac{1,96\times\sigma}{\sqrt{298}}; 18,4 + \frac{1,96\times\sigma}{\sqrt{298}}] = [18,0; 18,8]$. C'est l'intervalle de confiance au seuil 0,95 que l'on a vu dans le cours : l'observatoire a raison.
- Le collectif Jeudi Noir utilise un autre intervalle de confiance ! L'intervalle $I_2 = [18,1; +\infty[$ Cet intervalle est aussi un intervalle de confiance au seuil 0,95 (mais on ne l'a pas vu dans ce cours). Donc effectivement, le collectif a raison.
- Enfin, l'association des propriétaires-bailleurs utilise l'intervalle $I_3 =]-\infty; 18,7]$ (en fait$[0; 18,7]$ car un loyer ne peut pas être négatif). C'est aussi un intervalle de confiance au seuil 0,95 (mais on ne l'a pas vu dans ce cours). Donc l'association a raison.