

Échantillonnage

1 Principe de l'échantillonnage

La théorie de l'échantillonnage consiste à déterminer des propriétés sur des échantillons tirés au hasard parmi une population dont on connaît les propriétés.

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 Échantillonnage de variables aléatoires

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 note cet échantillon $\mathcal{E} = (\omega_1, \dots, \omega_n)$.

A chaque individu ω_i de cet échantillon, on associe une valeur: $x_i = X(\omega_i)$ et on note X_i la variable aléatoire qui à $\mathcal{E}$ associe x_i .

Exemple 1.

Ω =les étudiants de la classe;

X = cette variable aléatoire associe à chaque étudiant sa note au contrôle;

$\mathcal{E}$ =une suite de trois étudiants tirés au hasard;

$n=3$;

X_1 est la variable aléatoire qui donne la note de contrôle du premier étudiant de l'échantillon;

x_1 est la note du premier étudiant de l'échantillon;

Si l'échantillonnage est non-exhaustif (avec remise), les variables aléatoires X_i sont indépendantes et suivent la même loi que la variable aléatoire X .

On appelle statistique toute variable aléatoire Y qui est uniquement dépendante de $X_1, \dots, X_n$. Le but de l'échantillonnage est de déterminer la loi de la statistique Y en fonction de la loi de X qui est connue.

Exemple 2. On peut s'intéresser à la moyenne de X sur l'échantillon:

$$Y := M_n = \frac{X_1 + \dots + X_n}{n}.$$

C'est une statistique qui à chaque échantillon associe la moyenne de X sur l'échantillon (dans notre exemple précédent X représentait des notes).

Exemple 3. On peut aussi bien s'intéresser à la variance de X sur un échantillon:

$$Y := S_n^2 = \frac{(X_1 - \overline{X_n})^2 + \dots + (X_n - \overline{X_n})^2}{n}.$$

2.1 Quelques rappels de probabilités

Définition 1. Deux variables aléatoires X et Y sont dites indépendantes si pour tous événements A et B , on a:

$$\mathbb{P}(A \text{ et } B) = \mathbb{P}(A) \times \mathbb{P}(B).$$

Proposition 1. Si X et Y sont deux variables aléatoires et si a et λ sont des nombres réels, alors:

$$\begin{aligned} \mathbb{E}(X + a) &= \mathbb{E}(X) + a & V(X + a) &= V(X) \\ \mathbb{E}(\lambda X) &= \lambda \mathbb{E}(X) & V(\lambda X) &= \lambda^2 V(X) \\ \mathbb{E}(X + Y) &= \mathbb{E}(X) + \mathbb{E}(Y) & V(X + Y) &= V(X) + 2COV(X, Y) + V(Y) \end{aligned}$$

Proposition 2. Si X et Y sont deux variables aléatoires indépendantes alors:

$$Cov(X, Y) = 0 \quad V(X + Y) = V(X) + V(Y).$$

2.2 Cas des lois normales

Théorème 1. Soit X une variable aléatoire de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et Y une variable aléatoire de loi $\mathcal{N}(\mu_2, \sigma_2^2)$. Si les variables X et Y sont indépendantes, alors la variable aléatoire $X + Y$ suit une loi $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Cette proposition est très importante dans le cas de l'échantillonnage. En effet, supposons que nous avons une variable aléatoire X suivant une loi normale $\mathcal{N}(\mu; \sigma^2)$, alors les variables aléatoires X_i sont indépendantes et suivent la même loi normale $\mathcal{N}(\mu; \sigma^2)$. En utilisant la proposition ci-dessus, on obtient facilement que la moyenne sur l'échantillon M_n suit une loi normale $\mathcal{N}(\mu; \frac{\sigma^2}{n})$.

2.3 Cas général

Si la variable aléatoire X suit une loi quelconque, on peut déterminer la loi de M_n lorsque n est "grand". Une version précise de cet énoncé est donnée par le théorème suivant.

Théorème 2. (Théorème de la limite centrale) Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes suivant toutes la même loi de moyenne μ et de variance σ^2 . Alors pour n suffisamment grand (disons $n \geq 30$), la variable aléatoire

$$M_n := \frac{X_1 + \dots + X_n}{n}$$

suit approximativement une loi normale $\mathcal{N}(\mu; \frac{\sigma^2}{n})$.

3 Distribution d'échantillonnage

Grâce aux résultats que nous venons d'énoncer, nous allons pouvoir étudier ci-dessous deux cas particuliers d'échantillonnage.

3.1 Distribution d'échantillonnage des moyennes

Considérons une population ayant une certaine propriété avec une moyenne μ et une variance σ^2 .

Soit M_n la variable aléatoire qui à tout échantillon aléatoire (prélevé avec remise) d'effectif n fixé, associe la moyenne de cet échantillon.

Alors en appliquant le théorème de la limite centrale, on remarque que pour n suffisamment grand ($n \geq 30$),

M_n suit approximativement la loi normale $\mathcal{N}(\mu; \frac{\sigma^2}{n})$.

Remarque 1. On a les remarques suivantes :

- On considère en générale que n est suffisamment grand pour utiliser les résultats précédents lorsque $n \geq 30$.
- Si le caractère étudié sur la population suit une distribution normale, on peut utiliser ces résultats même si n est petit.

Exemple 4. Les statistiques des notes obtenues au BAC STI 2006 en France sont:

Moyenne nationale: $\mu = 10,44$,

Écart-type: $\sigma = 1,46$.

Une classe de BTS comporte 35 élèves en 2006/2007, tous issus d'une BAC STI en 2006.

Quelle est la probabilité que la moyenne des notes de BAC de cette classe soit supérieure à 10?

3.2 Distribution d'échantillonnage des pourcentages

Considérons une population dont un pourcentage p d'éléments possède une certaine propriété.

Soit K_n la variable aléatoire qui, à tout échantillon aléatoire (prélevé avec remise) d'effectif n fixé, associe le nombre d'individus qui ont la propriété étudiée. On sait que K_n suit une loi binomiale $\mathcal{B}(n; p)$.

Alors en appliquant le théorème de la limite centrale, on remarque que pour n suffisamment grand,

K_n suit approximativement la loi normale $\mathcal{N}(np; np(1 - p))$.

On s'intéresse ensuite à $F_n = \frac{K_n}{n}$ la variable aléatoire qui, à tout échantillon aléatoire (prélevé avec remise) d'effectif n fixé, associe la fréquence avec laquelle les éléments de cet échantillon possèdent cette propriété.

Alors on déduit que pour n suffisamment grand,

F_n suit approximativement la loi normale $\mathcal{N}(p; \frac{p(1 - p)}{n})$.

Exemple 5. Une élection a eu lieu et un candidat a eu 40% des voix. On prélève un échantillon de 100 bulletins de vote.

Quelle est la probabilité que, dans l'échantillon, le candidat ait entre 35% et 45% des voix?

4 Une petite remarque technique

Lorsqu'on travaille concrètement avec des lois normales, on travaille souvent sur des tables de loi normale car cette dernière est très difficile à calculer. Or ces tables donnent des valeurs de la fonction de répartition de la loi normale $\mathcal{N}(0; 1)$.

Mais en pratique, nous avons affaire à des lois normales ayant des paramètres qui peuvent être bien différents. Il est alors important de savoir se ramener au cas $\mathcal{N}(0; 1)$.

Proposition 3. Si X suit une loi normale $\mathcal{N}(\mu; \sigma^2)$, alors la variable aléatoire

$$\frac{X - \mu}{\sigma}$$

suit une loi $\mathcal{N}(0; 1)$.