Ch 5: Échantillonnage et estimation des paramètres

1. Échantillon, paramètre et statistique

1.1 Échantillon aléatoire

Un échantillon aléatoire est une suite de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi qu'une caractéristique X d'une population.

1.2 Paramètre

Un paramètre est un nombre qui décrit une caractéristique de la population étudiée. Citons, à titre d'exemples, la moyenne μ, la variance σ^2, la médiane M et la proportion p d'une population. Notons que les paramètres sont souvent inconnus.

1.3 Statistique

Une statistique est une fonction de l'échantillon qui permet d'estimer un paramètre de la population.

Par exemple :

a) La moyenne μ d'une population est estimée par la moyenne $\overline{X}$ d'un échantillon de cette population :

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

b) La variance σ^2 d'une population est estimée par la variance S^2 d'un échantillon de cette population :

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

c) Soit une population ayant une caractéristique qualitative (une maladie particulière). La proportion p des individus ayant cette caractéristique dans la population est estimée par :

$$\hat{p} = \frac{S_n}{n}$$

où S_n désigne le nombre d'individus de l'échantillon qui possèdent cette caractéristique.

Remarque

Notons que les statistiques $\overline{X}$, S^2 et $\hat{p}$ sont appelées aussi estimateurs. Par contre, la valeur calculée par un estimateur pour un échantillon donnée est appelée estimation ponctuelle.

2. Qualité d'un estimateur

2.1 Estimateur sans biais

Définition : Un estimateur T d'un paramètre θ est dit sans biais si son espérance mathématique est égale à la vraie valeur du paramètre à estimer :

$$E(T) = \theta$$

Notons qu'un estimateur sans biais ne surestime ni sous-estime systématiquement le paramètre. On dit d'un estimateur sans biais qu'il est bien centré.

Remarque : Notons que $\overline{X}$, S^2 et $\hat{p}$ sont respectivement des estimateurs sans biais des paramètres μ , σ^2 et p c'est-à-dire :

$$E(\overline{X}) = \mu, \quad E(S^2) = \sigma^2 \quad \text{et} \quad E(\hat{p}) = p.$$

2.2 Estimateur efficace

Définition : Soient T_1et T_2 deux estimateurs sans biais d'un paramètre inconnu θ. On dit que T_1 est plus efficace que T_2 si

$$Var(T_1) < Var(T_2)$$

Un estimateur sans biais doit avoir une variance aussi petite que possible, afin d'être aussi précis que possible. Ainsi les variances des estimateurs

$$Var(\overline{X}) = \frac{\sigma^2}{n} \quad \text{et} \quad Var(\hat{p}) = \frac{p(1-p)}{n}$$

Ces formules montrent que les variances de $\overline{X}$ et celle de $\hat{p}$ diminuent lorsque la taille n de l'échantillon augmente. Donc, plus l'échantillon est grand, plus $\overline{X}$ et $\hat{p}$ sont précis.

3. Distribution d'échantillonnage

Une statistique est par définition basée sur un échantillon qui n'est qu'une partie de la population étudiée; il est donc fort improbable que la valeur prise par cette statistique coïncide avec le paramètre étudié.

Définition : La distribution d'échantillonnage d'une statistique est la distribution de toutes les valeurs possibles de cette statistique. Ces valeurs sont calculées à partir de tout les échantillons de même taille et issus de la même population.

3.1 Étude de la distribution échantillonnale de $\overline{X}$

a. *Population normale de variance connue :*

Si $X_1,\ldots,X_n$ un échantillon issu d'une population de loi normale de variance σ^2 connue, alors, $\overline{X}$ suit une lois normale:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$$

b. Population normale de variance inconnue :

Si $X_1,\ldots,X_n$ un échantillon issu d'une population de loi normale de variance σ^2 inconnue, alors:

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \approx t_{n-1}$$

où t_{n-1} désigne la loi de Student de d.d.l $n-1$.

c. Population de loi inconnue :

Si $X_1, \ldots, X_n$ un échantillon issu d'une population de loi inconnue, alors le théorème central limite nous permet d'écrire :

$$Z = \frac{\overline{X} - \mu}{S / \sqrt{n}} \approx N(0,1)$$

pourvue que la taille de l'échantillon n soit assez grande ($n \geq 30$).

3.2 Étude de la distribution échantillonnale de la variance.

a. *Population normale de moyenne connue :*

Si $X_1, \ldots, X_n$ un échantillon aléatoire issu d'une population de loi normale de moyenne μ connue, alors,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \approx \chi_n^2,$$

la lois de Khi-carré à n d.l.l:

b. Population normale de moyenne inconnue :

Si $X_1, \ldots, X_n$ un échantillon aléatoire issu d'une population de loi normale de moyenne μ inconnue, alors,

$$\frac{n-1}{\sigma^2} S^2 \approx \chi^2_{n-1},$$

la lois de Khi-carré à n-1 d.l.l.

4. Estimation par intervalle de confiance.

4.1 Estimation par intervalle de confiance de la moyenne μ.

La moyenne $\overline{X}$ calculée à partir d'un échantillon donné est presque toujours un peu plus grande ou un peu plus petite que la vraie moyenne de la population μ. On cherche plutôt une approximation qui tient compte de la marge d'erreur d'estimation. Cette estimation se présente alors sous la forme :

$$\overline{X} \pm E$$

La marge d'erreur E est appelée précision de l'estimateur $\overline{X}$.

Ainsi l'estimation par intervalle de confiance de μ consiste à déterminer l'erreur E de façon que

$$\mu \in \left[\overline{X} - E, \overline{X} + E\right]$$

avec une probabilité égale à $1-\alpha$ appelée niveau de confiance.

Par exemple, on peut déterminer un intervalle de confiance qui contient la valeur de μ avec un niveau de confiance égal à 95%. Cela veut dire que si on répète la même procédure d'estimation 100 fois, la moyenne sera dans 95 intervalles parmi les 100 intervalles établis. Cela signifie que si on construit un intervalle de confiance par un seul échantillon, il y aura un risque de 5% que la valeur de μ ne sera pas dans cet intervalle.

Pour construire de tels intervalles de confiance, nous aurons besoin des quantiles de la loi normale et de la loi de Student définis ci-après.

4.1.1 Quantile d'ordre α des lois normale et Student

Fixons un nombre α dans l'intervalle $[0,1]$ et notons z_α **et** $t_{\alpha,v}$ les quantiles de la loi normale et de la loi de Student définis par :

z_α est la valeur telle que $P(Z \geq z_\alpha) = \alpha$

$t_{\alpha,v}$ est la valeur telle que $P(T \geq t_{\alpha,v}) = \alpha$

Exemple : Calculez à l'aide des tables: $z_{0.05}, z_{0.025},$

$t_{0.05,10}$ et $t_{0.025,12}.$

Tableau de certaines valeurs critiques de la loi normale:

$$\phi(z_\alpha) = 1 - \alpha$$

α	0.005	0.01	0.025	0.05	0.01
z_α	2.575	2.325	1.96	1.645	1.285
$z_{\alpha/2}$	2.807	2.575	2.241	1.96	1.645

4.1.2 Construction de l'intervalle de confiance de μ.

Soit $\left[\overline{X} - E, \overline{X} + E\right]$ un intervalle de confiance de μ.

Afin de déterminer la précision E, on distingue

quatre cas :

- Cas 1: Si $X_1,\ldots,X_n$ est un échantillon issu d'une population de loi normale de <u>variance</u> σ^2 <u>connue</u>, alors l'intervalle de confiance de niveau $1-\alpha$ de μ est :

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Ainsi la précision de l'estimation sera :

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Cas 2: Si $X_1,\ldots,X_n$ est un échantillon issu d'une population de loi normale de variance σ^2 inconnue, alors l'intervalle de confiance de niveau $1-\alpha$ de μ est :

$$\left[\overline{X} - t_{\alpha/2;n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2;n-1}\frac{S}{\sqrt{n}}\right]$$

La précision de l'estimation sera :

$$E = t_{\alpha/2;n-1}\frac{S}{\sqrt{n}}$$

- Cas 3: Si $X_1, \ldots, X_n$ est un échantillon issu d'une population de loi inconnue, alors pourvue que la taille n soit assez grande ($n \geq 30$), l'intervalle de confiance de niveau $1-\alpha$ de μ est :

$$\left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

De même la précision de l'estimation sera :

$$E = z_{\alpha/2}\frac{S}{\sqrt{n}}$$

- Cas 4: Si $X_1,\ldots,X_n$ est un échantillon choisi sans remise à partir d'une population de taille finie N et de loi inconnue, alors pourvue que la taille n soit assez grande ($n \geq 30$), l'intervalle de confiance de niveau $1-\alpha$ est :

$$\left[\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

Dans cette situation la précision sera :

$$E = z_{\alpha/2}\frac{S}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

4.2. Intervalle de confiance d'une proportion

Soit p la proportion d'individus dans la population ayant une caractéristique qualitative donnée.

4.2. 1 Intervalle de confiance d'une proportion pour une population infinie

L'intervalle de confiance de niveau $1-\alpha$ de p est de la forme :

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

pourvue que la taille n soit assez grande et

que $np \geq 5$ et $n(1-p) \geq 5.$

4.2.2 Intervalle de confiance d'une proportion pour une population finie, avec tirage sans remise

L'intervalle de confiance de niveau $1-\alpha$ de p est:

$$\left[\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}},\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

pourvue que $np \geq 5$ et $n(1-p) \geq 5.$

5. Choix de la taille d'échantillon

La qualité d'un intervalle de confiance se mesure par son degré de confiance $1-\alpha$ et sa marge d'erreur $E.$ Un choix adéquat de la taille de l'échantillon permet de contrôler simultanément ces deux facteurs.

5.1 Cas d'une moyenne μ.

Dans le cas d'une population normale de variance connue (cas 1), nous pouvons déterminer la taille minimale requise de l'échantillon pour avoir un intervalle de confiance de niveau $1-\alpha$ au moins et de précision fixés e à l'avance:

$$n \geq \left[\frac{z_{\alpha/2}}{e} \sigma \right]^2$$

Lorsque σ est inconnu, on le remplacera par une pré-estimation S.

5.2 Cas d'une proportion p.

Dans le cas d'une proportion, si l'on dispose d'une pré-estimation $\hat{p}$ de p , la taille minimale sera :

$$n \geq \left[\frac{z_{\alpha/2}}{e} \right]^2 \hat{p}(1-\hat{p})$$

Par contre, si la pré-estimation n'est pas disponible, la taille requise sera alors :

$$n \geq \left[\frac{z_{\alpha/2}}{2e} \right]^2$$

Remarque : Si la population est de taille finie N et le tirage est sans remise, alors :

1) La taille requise pour la moyenne sera :

$$n \geq \frac{N z_{\alpha/2}^2 \sigma^2}{(N-1)e^2 + z_{\alpha/2}^2 \sigma^2}$$

Ainsi si la variance est inconnue, on la remplace par une pré-estimation.

2) La taille requise pour la proportion sera :

$$n \geq \frac{N z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{(N-1)e^2 + z_{\alpha/2}^2 \hat{p}(1-\hat{p})}$$

Lorsqu'on ne possède pas de pré-estimation $\hat{p}$, on prendra:

$$\hat{p} = 0.5$$