4 Théorèmes limites

(Convergence)

Remarques:
Dans une étude statistique, on fait un échantillonnage par lequel on détermine les paramètres statistiques :
* moyenne de X
* espérance de X
etc...
Le TCL nous explique comment on peut faire confiance à ces estimations:
* moyenne m
* espérance sigma

(Multivariables)

Théorèmes limites

4.1 Loi des grands nombres

Au vu des chapitres précédents, on peut énoncer le théorème suivant.

Théorème 49

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = nm$$

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$$

Définition 51 *La moyenne empirique des v.a. $X_1, \ldots, X_n$ est la v.a.*

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

(Moyenne des var)

On sait d'ores et déjà que la moyenne empirique a pour espérance m et pour variance σ^2/n. Ainsi, plus n est grand, moins cette v.a. varie. À la limite, quand n tend vers l'infini, elle se concentre sur son espérance, m. C'est la loi des grands nombres.

Théorème 52 (Loi des grands nombres) *Quand n est grand, $\overline{X}_n$ est proche de m avec une forte probabilité. Autrement dit,*

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|\overline{X}_n - m| > \varepsilon) = 0$$

preuve : soit $\varepsilon > 0$, d'après l'inégalité de Chebyshev,

$$P(|\overline{X}_n - m| > \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \longrightarrow 0$$

On dit que $\overline{X}_n$ converge en probabilité vers m.

Exemple 53 (Fréquence empirique et probabilité) *Il est ici question de rapprocher la probabilité théorique d'un événement A parfois inconnue avec la fréquence empirique de cet événement obtenue en répétant plusieurs fois (mettons n fois) l'expérience aléatoire qui produit A dans les mêmes conditions.*

4.2 Théorème central limite

Théorème 54 *Pour tous réels $a < b$, quand n tend vers $+\infty$,*

$$P\left(a \leq \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leq b\right) \longrightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x$$

On dit que $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}$ converge en loi vers la loi normale $\mathcal{N}(0,1)$. On a le conséquence (presque) immédiate :

Considérons une suite $(X_n)_{n\geq 1}$ de v.a. indépendantes et de même loi. Supposons que ces v.a. ont une espérance, notée m et une variance, notée σ^2.

Corollaire 55 *Quand n est grand, la loi binomiale $\mathcal{B}(n,p)$ est proche de la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$*

4.3 Intervalles de confiance

Soit X un caractère (ou variable) étudié sur une population, de moyenne m et de variance σ^2. On cherche ici à donner une estimation de la moyenne m de ce caractère, calculée à partir de valeurs observées sur un échantillon $(X_1, ..., X_n)$.

La fonction de l'échantillon qui estimera un paramètre est appelée **estimateur**, son écart-type est appelé **erreur standard** et est noté SE. L'estimateur de la moyenne m est la moyenne empirique

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

Estimateur de la moyenne réelle

Que savons-nous sur $\overline{X}_n$?
1) $E[\overline{X}_n] = m$,
2) $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$, donc $\mathrm{SE}(\overline{X}_n) = \sigma/\sqrt{n}$,
3) $\overline{X}_n \longrightarrow m$ quand $n \to \infty$,
4) quand n est grand, $\overline{X}_n$ suit approximativement une loi normale $\mathcal{N}(m, \sigma/\sqrt{n})$.

D'après les propriétés de la loi normale, quand n est grand (mettons supérieur à 20), on sait que

$$P[m - 2\sigma/\sqrt{n} \le \overline{X}_n \le m + 2\sigma/\sqrt{n}] = 0.954$$

ou, de manière équivalente,

$$P[\overline{X}_n - 2\sigma/\sqrt{n} \le m \le \overline{X}_n + 2\sigma/\sqrt{n}] = 0.954$$

Ce qui peut se traduire ainsi : quand on estime m par $\overline{X}_n$, l'erreur faite est inférieure à $2\sigma/\sqrt{n}$, pour 95,4% des échantillons. Ou avec une probabilité de 95,4%, la moyenne inconnue m est dans l'intervalle $[\overline{X}_n - 2\sigma/\sqrt{n}, \overline{X}_n + 2\sigma/\sqrt{n}]$. IDC pour la moyenne (1)

Définition 56 *On peut associer à chaque incertitude α, un intervalle, appelé intervalle de confiance de niveau de confiance $1-\alpha$, qui contient la vraie moyenne m avec une probabilité égale à $1-\alpha$.*

Définition 57 *Soit Z une v.a.. Le fractile supérieur d'ordre α de la loi de Z est le réel z qui vérifie*

$$P[Z \ge z] = \alpha$$

Le fractile inférieur d'ordre α de la loi de Z est le réel z qui vérifie

$$P[Z \le z] = \alpha$$

Proposition 58 *Un intervalle de confiance pour la moyenne, de niveau de confiance $1-\alpha$, est*

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

IDC pour la moyenne (2)

où $z_{\alpha/2}$ est le **fractile** *supérieur d'ordre $\alpha/2$ de la loi normale $\mathcal{N}(0,1)$.*

Remarque : soit Z une v.a. de loi $\mathcal{N}(0,1)$, $z_{\alpha/2}$ vérifie

$$P[Z \le -z_{\alpha/2}] = P[Z \ge z_{\alpha/2}] = \frac{\alpha}{2}, \qquad P[-z_{\alpha/2} \le Z \le z_{\alpha/2}] = 1 - \alpha$$

On appelle aussi intervalle de confiance la réalisation de l'intervalle précédent

$$[\bar{x}_n - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x}_n + z_{\alpha/2}\sigma/\sqrt{n}]$$

Seules quelques valeurs de α sont utilisées habituellement. Les trois valeurs communes sont :
- $\alpha = 0.01$, et $z_{0.005} = 2.58$,
- $\alpha = 0.05$, et $z_{0.025} = 1.96$,
- $\alpha = 0.1$, et $z_{0.05} = 1.645$.

Quand l'écart-type théorique de la loi du caractère X étudié n'est pas connu, on l'estime par l'écart-type empirique s_{n-1}. Comme on dispose d'un grand échantillon (de taille supérieure à 20), l'erreur commise est petite. L'intervalle de confiance, de niveau de confiance $1-\alpha$ devient :

$$\left[\bar{x}_n - z_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}},\ \bar{x}_n + z_{\alpha/2}\frac{s_{n-1}}{\sqrt{n}}\right]$$

où

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Exemple 60 (suite) On peut maintenant donner un intervalle de confiance, de niveau de confiance 95%, pour le volume sonore moyen le long de la RN 7. Calculons :

$$\overline{x} = 54.6552, \qquad s_{n-1} = 2.9553$$

d'où l'intervalle

$$\left[54.6552 - \frac{1.96 * 2.9553}{\sqrt{29}}, 54.6552 + \frac{1.96 * 2.9553}{\sqrt{29}}\right] = [53.5796, 55.7308]$$

Le bruit moyen occasionné par le trafic routier de la nationale 7 est compris entre 53,58 et 55,73, avec un niveau de confiance de 95%.

Tout ce qui concerne les moyennes s'applique aussi aux proportions. On considère une caractéristique que possède une proportion p d'individus dans la population. Soit un échantillon de taille n pris dans la population et X le nombre d'individus dans cet échantillon qui possède la caractéristique étudiée. Alors

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

est un estimateur de p. De plus, on a déjà vu que X suit une loi binomiale $\mathcal{B}(n,p)$ et que cette loi peut être approchée par une loi normale :

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

suit approximativement une loi normale $\mathcal{N}(0,1)$.
On peut estimer l'écart-type de $\hat{p}$ qui est $\sqrt{p(1-p)/n}$, par $\sqrt{\hat{p}(1-\hat{p})/n}$. On obtient alors un intervalle de confiance pour la proportion p inconnue, de niveau de confiance $1-\alpha$

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

IDC pour la proportion

Rappel : cet intervalle n'est valide que lorsque $n \geq 20$.

4.4 Exercices

Exercice 1 – Une enquête marketing a pour but de vérifier si les cibles potentielles seraient tentées par un nouveau produit. Il a été montré que 56% des gens sont favorables au nouveau produit. Pour aller plus loin, on interroge à nouveau 200 personnes. Quelle est la loi du nombre de clients potentiels parmi les 200 ? Par quelle loi peut-on l'approcher ? Calculer $P[X > 100]$ et $P[100 \leq X \leq 150]$.

On souhaite maintenant demander des précisions à un grand nombre de personnes favorables au produit, mettons 100 personnes. Déterminer la taille de l'échantillon de personnes à interroger pour que notre échantillon contienne au moins 100 personnes favorables, avec une probabilité supérieure ou égale à 95%.

Exercice 2 – On prélève indépendamment et avec remise n individus d'une population séparée en deux sous-populations A et A^c de proportions respectives p et $1-p$ (clients importants ou petits clients par exemple).

1) Soit K le nombre d'individus de la sous-population A présents dans l'échantillon. Quelle est la loi de K ?

2) Notons $F = K/n$ la fréquence empirique de la catégorie A. Donner l'espérance et la variance de F.

3) Quel est le comportement de F quand n devient grand ?

4) On considère un échantillon de 400 clients d'un organisme de placement financier. Les clients de cet organisme, dans leur ensemble, se répartissent ainsi : 20% possèdent de gros portefeuilles, les autres ne détenant que des portefeuilles modestes. Quelle est la probabilité que la proportion F de gros clients dans l'échantillon soit comprise entre 7% et 23% ? Quelle est la probabilité pour que F soit inférieure à 15% ? Dans l'agence de Lyon (qui compte 400 clients), seuls 15% des clients sont de gros clients. Est-ce acceptable par le siège ?

Exercice 3 – On suppose que le poids d'un nouveau né est une variable normale d'écart-type égal à 0,5 kg. Le poids moyen des 49 enfants nés au mois de janvier 2004 dans l'hôpital de Charleville-Mézières a été de 3,6 kg.

a) Déterminer un intervalle de confiance à 95% pour le poids moyen d'un nouveau né dans cet hôpital.

b) Quel serait le niveau de confiance d'un intervalle de longueur 0,1 kg centré en 3,6 pour ce poids moyen ?

Exercice 4 – On veut étudier la proportion p de gens qui vont au cinéma chaque mois. On prend donc un échantillon de taille $n = 100$. Soit N le nombre de personnes dans l'échantillon qui vont au cinéma mensuellement.

1) Quelle est la loi de N ? Par quelle loi peut-on l'approcher et pourquoi ? En déduire une approximation de la loi de $F = N/n$.

2) On observe une proportion f de gens qui vont chaque mois au cinéma. Donner la forme d'un intervalle de confiance pour p, de niveau de confiance $1 - \alpha$.

3) Applications numériques : $f = 0,1$, $1 - \alpha = 90\%,\ 95\%,\ 98\%$.