Chapitre 5

ESTIMATION ET INTERVALLES DE CONFIANCE

1. DES PROBABILITÉS À LA STATISTIQUE.

hypothèse intuitive élaborée à partir d'expériences diverses : convergence de la proportion dans laquelle un événement est réalisé au cours d'expériences répétées vers sa probabilité telle que nous l'avons définie dans le chapitre précédent.

1.1 Simulation.

Exemple de tableau de données simulées (ou table de nombres au hasard) :

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	*0.833*	*0.275*	*0.972*	*0.004*	*0.978*	*0.532*	*0.376*	*0.516*
2	*0.518*	*0.936*	*0.341*	*0.333*	*0.177*	*0.879*	*0.010*	*0.090*
3	*0.863*	*0.195*	*0.187*	*0.439*	*0.436*	*0.870*	*0.226*	*0.374*

Tableau 1.5 *: nombres pseudo-aléatoires*

Pour obtenir des nombres compris entre –1 et 2 on effectue la transformation suivante, pour toute valeur x du tableau précédent :

$$y = 3x - 1$$

Pour obtenir des nombres entiers compris entre 1 et 6, on pose :

$$Y = Int(6x + 1)$$

Int(y) désignant le plus grand entier inférieur ou égal à y : Int(5.456) = 5, Int(4) = 4.

1.2 Loi des grands nombres.

Cas d'un dé à 6 faces parfaitement équilibré : $\mathcal{P} = \{1, 2, 3, 4, 5, 6\}$

- à chaque jet, la probabilité d'obtenir {1} est égale à 1/6, et la face obtenue au i^e jet n'a aucune incidence sur les autres faces obtenues : il y a équiprobabilité, et les lancers sont indépendants.
- l'expérience montre que, pour n suffisamment grand, la proportion de faces {1} va tourner autour de 1/6. De même la proportion de faces {2}, de faces {3} etc.
- considérons l'événement A = {1, 2, 3, 4} : P(A) = 4/6 = 2/3. L'événement A se produit dans une proportion égale à la somme des proportions de chaque face et est donc de l'ordre de 4 x 1/6 soit 2/3. Cette proportion est là aussi de l'ordre de la probabilité.

Axiome de la loi des grands nombres : On considère une population contenant N unités statistiques. On y effectue n tirages avec remise et on compte le nombre n_A de réalisations d'un événement A donné d'effectif N_A. La proportion observée n_A/n converge vers la probabilité N_A/N de l'événement A lorsque le nombre de tirages augmente indéfiniment.

1.3 Notion de convergence.

La loi des grands nombres indique qu'il y a convergence des proportions vers les

probabilités, mais cette convergence dépend elle-même du hasard. Elle n'est pas systématique.

Définition : la convergence de la proportion de réalisations d'un événement au cours d'une suite d'expériences vers sa probabilité est appelée « convergence en probabilité ».

Propriété fondamentale : la densité observée d'une variable qualitative ou discrète converge en probabilité vers la densité de la v.a. lorsque le nombre d'observations augmente indéfiniment.

1.4 Densité et histogramme.

Propriété fondamentale : la densité observée d'une variable quantitative converge « en probabilité » vers la densité de la v.a. lorsque le nombre d'observations augmente indéfiniment et que la longueur des intervalles tend vers 0.

2. ESTIMATEUR D'UN PARAMÈTRE.

2.1 Estimations empiriques.

Définition : On appelle estimation empirique de la moyenne d'une variable aléatoire la moyenne calculée sur les observations effectuées.

On appelle estimation empirique de la variance d'une variable aléatoire la variance calculée sur les observations effectuées.

Propriété : Les estimations empiriques de la moyenne et de la variance convergent en probabilité vers les paramètres théoriques lorsque le nombre d'observations augmente indéfiniment.

2.2 Estimateurs de la moyenne et de la variance.

Définitions :

- L'échantillon de v.a. X_i, i = 1, …, n, est une suite de v.a. indépendantes et de même loi que X, la v.a. X_i représentant simplement la v.a. X au $i^{ième}$ tirage.
- L'échantillon observé x_i, i = 1, …, n, est une suite de valeurs observées de la v.a. X ou de chaque v.a. X_i, i = 1, …, n.

Définition :

On appelle estimateur d'un paramètre d'une loi de probabilité d'une v.a. X une v.a. calculée sur un échantillon X_i, i = 1, …, n de X, dont la valeur observée est une approximation de ce paramètre, et qui vérifie certaines propriétés d'optimalité.

Définitions :

- L'estimateur empirique de la moyenne théorique d'une v.a. est la v.a. M :

$$M = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- L'estimateur empirique de la variance théorique est la v.a. S^2 :

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - M)^2$$

2.3 Propriétés caractéristiques des estimateurs.

Un estimateur d'un paramètre ω est :

- sans biais si son espérance est égale à ω ;
- asymptotiquement sans biais si son espérance converge vers ω lorsque le nombre d'observations tend vers l'infini ;
- convergent si sa valeur observée converge en probabilité vers ω lorsque le nombre d'observations tend vers l'infini ;
- efficace s'il n'existe pas d'estimateur sans biais de ω de variance strictement inférieure.

Les estimateurs empiriques précédents possèdent des propriétés particulières :

- L'estimateur empirique de la moyenne est sans biais.
- L'estimateur empirique de la variance est asymptotiquement sans biais.
- Ils sont convergents.
- Lorsque les v.a. X_i suivent la loi normale, l'estimateur empirique de la moyenne est efficace.

2.4 Distributions d'échantillonnage.

Théorème de la limite centrée : on considère une suite de n v.a. X_i indépendantes et de même loi de probabilité, d'espérance μ et de variance σ^2. La loi de probabilité de l'estimateur M est, pour une valeur suffisante de n, la loi normale d'espérance μ et de variance σ^2/n.

Exemple *:*

On lance 100 fois le dé. La moyenne empirique M définie par la moyenne des 100 chiffres obtenus suit approximativement la loi normale d'espérance $\mu = 3.5$ et de variance $\sigma^2/n = 0.0292$:

- *La probabilité de l'intervalle $[\mu - 1.96\ \sigma/\sqrt{n},\ \mu + 1.96\ \sigma/\sqrt{n}] = [3.336, 3.664]$ est égale à 0.95. Il est donc très probable que la valeur moyenne obtenue en lançant le dé 100 fois soit comprise entre ces deux valeurs.*
- *La probabilité de l'intervalle $]-\infty\ ,\ \mu - 1.6449\ \sigma/\sqrt{n}] =]-\infty\ ,\ 3.219\ ]$ est égale à 0.05. On est presque sûr d'obtenir une valeur moyenne supérieure à 3.219.*

2.5 Loi de l'estimateur de la variance.

Théorème : si les v.a. X_i , i= 1, …, n sont indépendantes et suivent la loi normale d'espérance μ et de variance σ^2, la v.a. $n\ S^2/\sigma^2$ suit la loi du χ^2 de degré de liberté n – 1.

Exemple *: supposons n = 50 et $\sigma^2 = 25$. La v.a. $X^2 = 50\ S^2/25 = 2\ S^2$ suit la loi du χ^2 de degré de liberté 49. La table statistique pour $\nu = 49$ degrés de liberté donne les valeurs suivantes :*

$$P(2S^2<31.555) = 0.025 \quad P(2S^2>70.722) = 0.975$$

On en déduit la probabilité ci-dessous :

$$P(31.555<2S^2<70.722) = 0.95$$

La variance de l'échantillon sera très probablement comprise entre 15.778 et 35.361 :

$$\boxed{P(15.778 < S^2 < 35.361) = 0.95}$$

3. ESTIMATION PAR INTERVALLE DE CONFIANCE.

3.1 Généralités.

Définition : L'intervalle de confiance d'un paramètre d'une loi de probabilité est un intervalle observé sur un échantillon de cette loi contenant vraisemblablement la vraie valeur du paramètre.

Pour construire cet intervalle de confiance, on définit deux variables aléatoires B_1 et B_2 telles que la probabilité que l'intervalle [B1, B2] contienne la vraie valeur du paramètre soit élevée et égale à $1 - \alpha$ (par exemple, $1 - \alpha = 0.95$).

L'intervalle de confiance est donc la réalisation de cet intervalle aléatoire.

Définition : Le niveau de confiance est la probabilité que l'intervalle aléatoire contienne la vraie valeur du paramètre.

3.2 Intervalle de confiance de la moyenne (variance inconnue).

Définition : lorsque la variance théorique σ^2 est inconnue et estimée par s^2, l'intervalle de confiance de la moyenne au niveau de confiance $(100-\alpha)\%$ est l'intervalle :

$$\boxed{[m - t_\alpha\ s/(n-1)^{1/2}\ ,\ m + t_\alpha\ s/(n-1)^{1/2}\ ]}$$

Pour déterminer t_α, on utilise :

- si $n \leq 120$, la table de la loi de probabilité de Student de degré de liberté $\nu = n-1$;
- pour $n > 120$ la table de la loi normale centrée réduite.

Nous donnons ci-dessous quelques valeurs de t_α :

n = 10	ν = 9	α= 5%	t_α = 2.26	n = 20	ν = 19	α= 10%	t_α = 1.73
n = 20	ν = 19	α= 5%	t_α = 2.09	n = 50	ν = 49	α= 5%	t_α = 2.01

Exemple *: nous avons calculé dans le chapitre 1 la moyenne et la variance des 50 achats de l'échantillon tiré au hasard : m = 316.945F, s = 207.1291, s^2 = 42902.472. On a, pour $\alpha = 5\%$, $t_\alpha = 2.02$.L'intervalle de confiance de la moyenne est égal à :*

$$[316.945 - 2.02 \times 207.1291/\sqrt{49}\ ,\ 316.945 + 2.02 \times 207.1291/\sqrt{49}\ ]$$

Soit :

$$[\ 257.173,\ 376.717\ ]$$

Dans le calcul de l'intervalle de confiance de la moyenne, le manque de symétrie de la

répartition, constaté précédemment par l'étude de l'histogramme et la valeur du coefficient d'asymétrie (1.16) est compensé par le nombre d'observations (50).

3.3 Intervalle de confiance de la variance.

Définition : l'intervalle de confiance de la variance au niveau de confiance $(100 - \alpha)\%$ est l'intervalle :

$$[n\ s^2/\chi_{1-\alpha}^2\ ,\ n\ s^2/\chi_{\alpha}^2\ ]$$

Pour obtenir un intervalle de probabilité $1 - \alpha$, il faut déterminer deux bornes :

- χ_{α}^2 telle que $P(n\ S^2/\sigma^2 < \chi_{\alpha}^2) = \alpha/2$
- $\chi_{1-\alpha}^2$ telle que $P(n\ S^2/\sigma^2 > \chi_{1-\alpha}^2) = \alpha/2$

Pour déterminer les bornes χ_{α}^2 $\chi_{1-\alpha}^2$, on utilise la table statistique de la loi de du χ^2 de degré de liberté $\nu = n-1$. On sait en effet que la statistique $n\ S^2/\sigma^2$ suit cette loi de probabilité.

Le calcul de l'intervalle de confiance de la variance est plus compliqué pour n>100 et nous n'en parlerons pas (la procédure est expliquée dans la plupart des tables du $\chi 2$).

Exemple *: calculons l'intervalle de confiance de la variance des achats des clients d'Euromarket. L'estimation est* $s^2 = 42902.472$*. Le degré de liberté est égal à 49 pour 50 observations. On a, en choisissant un niveau de confiance égal à 95% :*

$$\chi_{\alpha}^2 = 31.555 \quad \chi_{1-\alpha}^2 = 70.222$$

D'où l'intervalle de confiance de la variance des achats :

$$[50x\ 42\ 902.472/70.222\ ,\ 50\ x\ 42\ 902.472/31.555\ ]$$

$$IC = [30\ 547.74,\ 67\ 980.47]$$

On sait que le montant des achats n'est pas réparti suivant la loi normale dans la population. On accordera donc un intérêt limité à l'intervalle de confiance ci-dessus que nous n'avons calculé qu'à titre d'exemple numérique.