Cours de Statistiques

Focus sur le Chapitre 4: Estimations

Licence 3 – Parcours Gestion et Finance

Stéphane ROBIN
Université Paris 1 Panthéon-Sorbonne
Département de Gestion – EM Sorbonne

Estimation d'un paramètre

Il existe deux grandes approches pour estimer un paramètre θ :

1. Approche **non paramétrique**: on cherche la **forme** de la loi théorique (inconnue *a priori*) qui correspond le mieux à la distribution du caractère observé.
2. Approche **paramétrique**: on suppose connue la forme de la loi qui sous-tend la distribution du caractère, et on cherche les "meilleures" valeurs des **paramètres inconnus** de cette loi

Ce cours ne traite que de l'approche paramétrique de l'estimation. Un aperçu de l'approche non paramétrique sera donné dans le cadre du Chapitre 5 (test statistiques)

Estimation ponctuelle et estimation par intervalle

Quelle que soit l'approche retenue, on peut :

1. Chercher une valeur numérique pour le paramètre θ à estimer: c'est **l'estimation ponctuelle**, dont le principe général fait l'objet de la présente section.
2. Chercher une plage de valeurs numériques dans laquelle on peut situer θ : c'est l'estimation par **intervalle de confiance**, à laquelle est consacré la seconde section

Mais, que l'on cherche une valeur ponctuelle de θ ou un intervalle plausible de valeurs de θ, l'estimation (ponctuelle ou par intervalle) aura pour point de départ le même estimateur.

Contenu du Chapitre 4

1. Estimation ponctuelle
    1. Propriétés souhaitables des estimateurs
    2. La moyenne comme estimateur de l'espérance
    3. Méthode du Maximum de Vraisemblance
2. Estimation par intervalle de confiance
    1. I.C. pour une proportion
    2. IC pour l'espérance d'une loi normale
    3. IC pour la variance d'une loi normale

1. Estimation ponctuelle

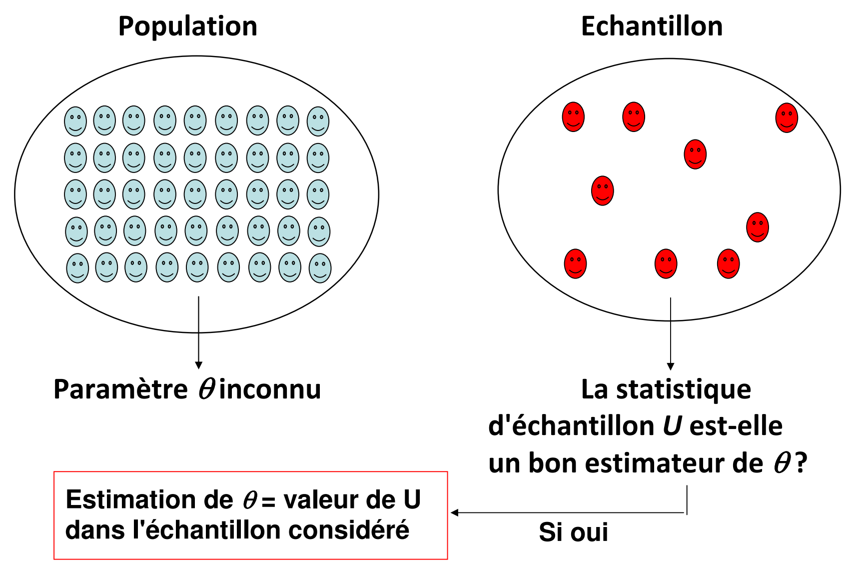
<u>Principe:</u> soit θ un paramètre inconnu (espérance par exemple) d'une population . On cherche la (les) statistique(s) d'échantillon qui représente(nt) le mieux le paramètre θ.

Cette statistique, qui permet de faire de l'inférence sur θ, est un **estimateur** de θ.

Une fois trouvé un bon estimateur de θ, on peut "faire une **estimation** de θ", c'est-à-dire attribuer à θ la valeur numérique de l'estimateur calculée dans un échantillon donné.

L'objectif de cette première section est de caractériser un "bon estimateur" de θ.

Principe de l'estimation ponctuelle



Questionnement

La distribution inconnue $p(X)$ d'un caractère X dans la population peut être résumée par ses moments (cf. Chap. 3): l'espérance m et la variance σ^2, qui ont chacune une valeur fixe mais inconnue.

Tout statistique d'échantillon (moyenne, par ex.) est une VA (cf. Chap. 3). Quelle statistique d'échantillon constitue un bon estimateur de l'espérance ? De la variance ?

<u>Ex. de question:</u> la moyenne "X barre" est-elle un bon estimateur de l'espérance ? Ou vaut-il mieux recourir à la médiane ?

Pour répondre à ce type de question, il faut préciser la notion d'estimateur et définir les **propriétés d'un bon estimateur**.

Notion d'estimateur

Dire que *U* est un **estimateur** de θ revient à dire que *U* est une fonction d'un échantillon de données tirées aléatoirement dans une population. *Bis repetita*: *U* est une VA (cf. Chap. 3)

Évalué dans plusieurs échantillons tirés de la même population, *U* peut prendre différentes valeurs. On dit que *U* fournit différentes **estimations** de θ, dont chacune est une réalisation de la VA *U*.

Dans cette section, on s'intéresse aux propriétés générales qui font de *U* un bon estimateur de θ, et non aux valeurs numériques que peut prendre *U* (c-à-d. aux diverses estimations de θ).

On verra dans la Section 2 comment, à partir d'une estimation ponctuelle de θ, trouver un intervalle de valeurs plausibles de θ.

1.1. Propriétés d'un bon estimateur

Un estimateur est un bon estimateur s'il est :

- **sans biais**
- le plus **efficace** possible
- **convergent**

Chacune de ces propriétés (absence de biais, efficacité, convergence) a une définition précise, dont on va donner dans ce qui suit la formulation la plus simple possible.

Absence de biais

Soit U un estimateur du paramètre θ. U est un estimateur sans biais de θ ssi:

$$\boxed{E(U) = \theta}$$

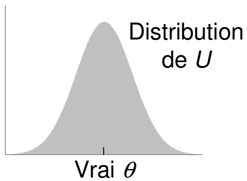
Autrement dit, U est sans biais s'il est "en moyenne" (sur plusieurs échantillons) égal à θ.

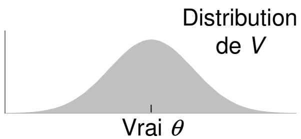
Un estimateur V de θ sera dit "biaisé" si $E(V) \neq \theta$. On définit ainsi le biais d'un estimateur V par la différence:

$$\boxed{\text{Biais de } V = E(V) - \theta}$$

Efficacité d'un estimateur sans biais

Un plus d'être sans biais, un bon estimateur doit avoir la variance plus faible possible, car sa valeur (aléatoire) sera plus souvent proche de la vraie valeur inconnue de θ.





Ici, U est plus efficace que V car Var U < Var V. L'**efficacité** de V comparée à celle de U est donnée par le rapport des variances:

$$\boxed{Efficacité = \frac{Var\,U}{Var\,V}}$$

Efficacité d'un estimateur quelconque

U, V et W sont trois estimateurs du paramètre θ. On observe que:

- U est sans biais, mais sa variance est très élevée
- V est faiblement biaisé, et sa variance est assez faible
- W est très fortement biaisé, mais il a la plus faible variance

Quel estimateur doit-on privilégier ? En général, celui qui a la plus faible dispersion autour de l'objectif θ, mesurée par:

$$E\,[(V-\theta)^2] = Var\,V + [\underbrace{E(V)-\theta}_{\text{Biais}}]^2$$

L'efficacité de V comparée à celle de U s'écrit maintenant:

$$\boxed{Efficacité = \frac{E\left[(U-\theta)^2\right]}{E\left[(V-\theta)^2\right]}}$$

Convergence d'un estimateur

Formellement, U est un estimateur convergent de θ si la probabilité que U soit dans un "petit intervalle" du vrai θ approche 1 quand la taille n de l'échantillon devient "grande":

$$Pr(\ |U - \theta| < \varepsilon\ ) = 1 \text{ quand } n > n_0$$

avec ε *réel* aussi petit que l'on veut et n_0 entier aussi grand que l'on veut (à la limite, $n \to \infty$).

Alternativement, U est un estimateur convergent de θ si la probabilité que U soit "loin" du vrai θ approche 0 quand la taille de l'échantillon devient "grande":

$$Pr(\ |U - \theta| > \eta\ ) = 0 \text{ quand } n > n_0$$

avec η réel et n_0 entier aussi grands que l'on veut.

Convergence: condition suffisante

Une des conditions qui rendent un estimateur U de θ convergent est que $[E(U - \theta)]^2 = 0$ quand $n \to \infty$ (n = taille de l'échantillon).

Comme $[E(U - \theta)]^2 = Var\ U + [E(U) - \theta]^2$, la condition ci-dessus peut être reformulée ainsi:

L'estimateur U de θ est convergent si sa variance et son biais tendent tous deux vers zéro quand la taille de l'échantillon devient grande (quand $n \to \infty$).

Cas des estimateurs asymptotiquement sans biais

On dit qu'un estimateur U de θ est **asymptotiquement sans biais** si son biais tend vers 0 quand la taille de l'échantillon augmente (quand $n \to \infty$).

Formellement: l'estimateur U_n calculé dans un échantillon de taille n est asymptotiquement sans biais si $\lim_{n \to \infty} |\ E(U_n) - \theta\ | = 0$.

Si la variance de U tend également vers 0 quand l'échantillon devient grand, alors U est un estimateur convergent de θ.

1.2. Exemple: estimateurs de l'espérance

Soit m la moyenne (espérance) inconnue dans une population.

On sait (cf. chap. 3) que si l'on tire au sort des échantillons de taille n dans cette population, la distribution d'échantillonnage de la moyenne empirique "X barre" est approximativement normale avec :

$$E(\overline{X}) = m$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

On va voir ci-après que ces propriétés font de "X barre" un bon estimateur de m.

Propriétés de la moyenne empirique (1)

1. Absence de biais
Par définition (cf. supra), la propriété $E(\overline{X}) = m$ implique que la moyenne empirique "X barre" est un estimateur sans biais de *m*.

2. Efficacité
Il peut exister plusieurs estimateurs concurrents de *m*. Par ex., la tendance centrale d'une population *symétrique* peut être estimée sans biais par la moyenne "X barre" et par la médiane Q_2.

Dans certains cas, la moyenne empirique est plus efficace que la médiane, et dans d'autres cas c'est le contraire (par ex., pour les distributions où il est probable de tirer des valeurs extrêmes).

Propriétés de la moyenne empirique (2)

2. Efficacité (suite)
Que si passe-t-il par ex. dans le cas d'une population normale? Dans ce cas, on a $Var(\overline{X}) = \sigma^2 / n$ et on démontre que Var(Q_2) = 1.57σ^2 / *n.*

Comme les deux estimateurs sont sans biais, l'efficacité de "X barre" comparée à celle de Q_2 est donnée par:

$$\frac{Var(Q_2)}{Var(\overline{X})} = \frac{1.57\sigma^2 / n}{\sigma^2 / n} = 1.57$$

La variance de Q_2 est de 57% plus élevée que celle de "X barre". Ici, la moyenne empirique est donc l'estimateur le plus efficace de l'espérance *m*.

Propriétés de la moyenne empirique (3)

3. Convergence
On a vu plus haut qu'un estimateur est convergent si son biais et sa variance tendent vers zéro quand la taille d'échantillon *n* est "grande".

Comme $E(\overline{X}) = m$, la moyenne empirique a un biais nul.

Comme $Var(\overline{X}) = \sigma^2 / n$, on a bien $\lim_{n \to \infty} Var(\overline{X}) = 0$

La moyenne empirique est donc bien un estimateur convergent de l'espérance de la population, *m*.

La moyenne empirique comme estimateur

On a montré que la moyenne empirique "X barre" est un bon estimateur de l'espérance de la population, *m*, car :
- "X barre" est un estimateur sans biais de *m*
- "X barre" est un estimateur efficace de *m*
- "X barre" est un estimateur convergent de *m*

Dès lors, pour obtenir une estimation ponctuelle de m, il suffit de calculer la valeur de "X barre" dans un échantillon donné.

Ainsi, dans le premier exemple du chap. 3 (âge des étudiants en L3 Finance), 21.8 et 22 sont deux estimations (dans des échantillons différents) de l'âge moyen *m* de la population.

1.3. Méthodes d'estimation

En statistique inférentielle, il existe plusieurs méthodes permettant de trouver un bon estimateur d'un paramètre, en particulier:

- Les Moindres Carrés Ordinaires (MCO)
- Le Maximum de Vraisemblance (MV)
- La Méthode des Moments

Elles ont des propriétés établies qui assurent que l'estimateur obtenu est (sous certaines conditions) sans biais, efficace et convergent. On utilise souvent le nom de la méthode pour désigner l'estimateur ainsi obtenu ("estimateur du MV", par ex.).

Dans le cadre de ce cours, nous allons examiner la méthode du **Maximum de Vraisemblance** et ses propriétés.

Vraisemblance d'un échantillon

Soit $\{X_1, X_2, \ldots, X_n\}$ un échantillon de n VA i.i.d., de même loi que le caractère X auquel on s'intéresse dans la population. Les valeurs $x_1, x_2, \ldots, x_n$ prises par $X_1, X_2, \ldots, X_n$ sont autant de réalisations possibles de la VA X (cf. chap. 3).

La probabilité d'observer, parmi tous les échantillons possibles de réalisations de X, les valeurs $x_1, x_2, \ldots, x_n$ est appelée **vraisemblance de l'échantillon**.

Elle est **conditionnelle** à la loi de X, caractérisée par au moins un paramètre inconnu θ (il peut y avoir plusieurs paramètres, comme dans le cas de la loi Normale). On la note $L(x_1, x_2, \ldots, x_n \mid \theta)$.

Formule de la vraisemblance

Les observations $x_1, x_2, \ldots, x_n$ étant indépendantes, la vraisemblance est simplement le produit des probabilités individuelles d'observer chaque valeur x_i :

$$L(x_1, x_2, \ldots, x_n \mid \theta) = L(x_1 \mid \theta)L(x_2 \mid \theta)\ldots L(x_n \mid \theta) = \prod_{i=1}^{n} L(x_i|\theta)$$

Plus précisément, dans le cas où X est un caractère continu, $L(x_i \mid \theta)$ représente la densité de probabilité de X évaluée au point x_i. Dans le cas où X est un caractère discret, $L(x_i \mid \theta) = Pr(X = x_i)$.

Les $x_1, x_2, \ldots, x_n$ étant connus et θ étant inconnu, $L(x_1, x_2, \ldots, x_n \mid \theta)$ peut être considérée comme une fonction de θ.

Le Maximum de Vraisemblance

La méthode du Maximum de Vraisemblance (MV) consiste, étant donné un échantillon de valeurs $x_1, x_2, \ldots, x_n$, à prendre comme estimateur de θ la valeur $\hat{\theta}$ de θ qui maximise la vraisemblance $L(x_1, x_2, \ldots, x_n \mid \theta)$. Par déf., $\hat{\theta}$ est solution de:

$$\frac{\partial L(x_1, x_2, \ldots, x_n|\theta)}{\partial \theta} = 0 \quad et \qquad \frac{\partial^2 L(x_1, x_2, \ldots, x_n|\theta)}{\partial \theta^2} \leq 0$$

ou, ce qui revient au même et se révèle plus pratique:

$$\frac{\partial \ln L(x_1, x_2, \ldots, x_n|\theta)}{\partial \theta} = 0 \qquad et \qquad \frac{\partial^2 \ln L(x_1, x_2, \ldots, x_n|\theta)}{\partial \theta^2} \leq 0$$

$\hat{\theta}$ est appelé "estimateur du MV" de θ. Sa valeur numérique dans l'échantillon considéré fournit une estimation de θ par la méthode du MV.

Propriétés de l'estimateur du MV

L'estimateur du MV a des propriétés utiles, résumées ci-dessous.

Théorème 1: sous certaines conditions de régularité de la fonction de vraisemblance, et si l'intervalle des valeurs de X ne dépend pas de θ, alors l'estimateur du MV $\hat{\theta}$

- converge en probabilité vers la vraie valeur du paramètre θ
- est asymptotiquement sans biais et efficace

Théorème 2: s'il existe un estimateur efficace de θ, alors il est solution de l'équation de vraisemblance $\partial L / \partial \theta = 0$.

Invariance fonctionnelle: si $\hat{\theta}$ est l'estimateur du MV de θ, alors $f(\hat{\theta})$ est l'estimateur du MV de $f(\theta)$

Vraisemblance et information

On appelle quantité d'information de Fisher $\boldsymbol{I_n(\theta)}$ apportée par un échantillon de taille n sur le paramètre θ la quantité suivante (positive ou nulle si elle existe):

$$I_n(\theta) = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]$$

Par théorème, si le domaine de définition du caractère X ne dépend pas de θ alors, si $\boldsymbol{I_n(\theta)}$ existe, on a:

$$I_n(\theta) = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

Information et efficacité d'un estimateur

Inégalité de Cramer-Rao: si le domaine de définition du caractère X ne dépend pas de θ et sous certaines conditions de régularité de la fonction de vraisemblance, la variance d'un estimateur sans biais $\hat{\theta}$ admet pour borne inférieure l'inverse de $\boldsymbol{I_n(\theta)}$:

$$Var(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$$

Théorème 3: un estimateur sans biais $\hat{\theta}$ est **efficace** si sa variance est **égale** à l'inverse de $\boldsymbol{I_n(\theta)}$

$$Var(\hat{\theta}) = \frac{1}{I_n(\theta)}$$

Cas de l'estimateur du MV

On a vu (Th. 1) que l'estimateur du MV est *asymptotiquement* sans biais (son espérance tend vers la vraie valeur du paramètre à estimer quand n, la taille de l'échantillon, tend vers l'infini).

On a vu également que, sous les conditions du Théorème 1, l'estimateur du MV $\hat{\theta}$ de θ est *asymptotiquement* efficace.

Cela signifie que $\lim_{n\to\infty} Var(\hat{\theta}) = \frac{1}{I_n(\theta)}$

On en déduit que si $\boldsymbol{I_n(\theta)}$ tend vers l'infini quand n tend vers l'infini, alors l'estimateur du MV est asymptotiquement convergent.

Méthode du MV pour plusieurs paramètres

La méthode du MV s'étend facilement à plus d'un paramètre. Dans le cas de k paramètres, la vraisemblance s'écrit:
$L(x_1, x_2, \ldots, x_n \mid \theta_1, \theta_2, \ldots, \theta_k)$.

Il faut alors, pour trouver les estimateurs des k paramètres, résoudre le système d'équations simultanées:

$$\frac{\partial \ln L}{\partial \theta_j} = 0 \qquad \text{avec } j = 1, 2, \ldots, k$$

Les propriétés de l'estimateur du MV rencontrées plus haut s'étendent au cas de plusieurs paramètres.

Quelques estimateurs du MV remarquables

L'estimateur du MV de l'espérance m d'une loi normale $\boldsymbol{N}(m, \sigma)$ est la moyenne empirique:

$$\hat{m} = \overline{X}$$

L'estimateur du MV de la variance σ^2 d'une loi normale $\boldsymbol{N}(m, \sigma)$ est :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Le paramètre m étant inconnu, on le remplace par son estimateur dans la formule de l'estimateur de la variance:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

Absence de biais

On peut vérifier aisément que $\hat{m}$ est un estimateur sans biais:

$$E(\hat{m}) = E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(x_i\right) = \frac{1}{n}\sum_{i=1}^{n} m = m$$

En revanche, $\hat{\sigma}^2$ est un estimateur biaisé. On peut montrer que:

$$E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

Cet estimateur est néanmoins asymptotiquement sans biais car:

$$\lim_{n\to\infty} \frac{n-1}{n} = 1 \quad \text{et donc} \quad \lim_{n\to\infty} E(\hat{\sigma}^2) = \sigma^2$$

Calcul de $E(\hat{\sigma}^2)$: remarque préliminaire

Ce calcul repose sur la **propriété** suivante: si $X_1, \ldots, X_n$ est une suite de VA i.i.d. de loi $\boldsymbol{N}(0,1)$, alors $\sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$

Dans le cas qui nous intéresse, on a $\forall i, \ \frac{X_i - m}{\sigma} \sim N(0,1)$

Et donc $\sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = \sum_{i=1}^{n}\frac{(X_i - m)^2}{\sigma^2} \sim \chi^2(n)$

Si on remplace m par son estimateur, cette relation devient:

$$\sum_{i=1}^{n}\frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

On ôte un degré de liberté car les x_i ne sont plus indépendants

Calcul de $E(\hat{\sigma}^2)$

Comme l'espérance d'une VA du χ^2 à p d.l. est égale à p, on a :

$$\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1) \Rightarrow E\left(\sum_{i=1}^{n} \frac{(x_i - \overline{X})^2}{\sigma^2}\right) = n-1$$

$$\Leftrightarrow \frac{1}{\sigma^2} E\left(\sum_{i=1}^{n} (x_i - \overline{X})^2\right) = n-1$$

$$\Leftrightarrow nE\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})^2\right) = \sigma^2(n-1)$$

$$\Leftrightarrow \boxed{E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2}$$

2. Estimation par intervalle de confiance

Principe: on s'intéresse à un paramètre inconnu d'une population (espérance par exemple).

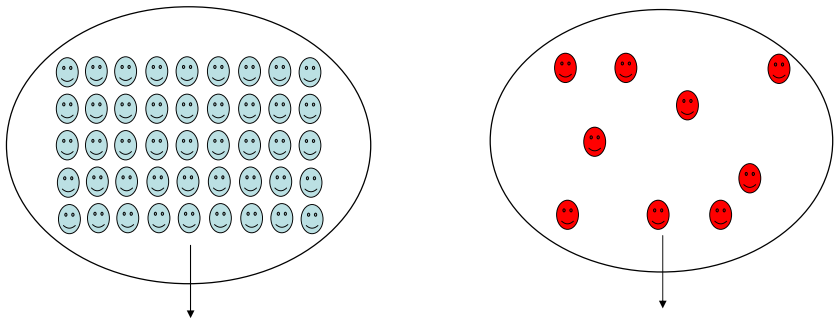
On calcule dans un échantillon une statistique correspondant à ce paramètre (moyenne pour l'espérance, par exemple). Cette statistique est appelée "estimateur" du paramètre.

La valeur de la statistique calculée dans l'échantillon (l'estimation) correspond-elle à la "vraie" valeur inconnue du paramètre?

Méthode: on choisit un "niveau de confiance" exprimé en % et on calcule à l'aide de l'estimation un encadrement du paramètre pour ce niveau de confiance. C'est "l'intervalle de confiance".

Principe de l'intervalle de confiance

Population **Echantillon**



Paramètre inconnu **Estimateur (statistique d'échantillon)**

$\boxed{a \leq \textbf{paramètre} \leq b \textbf{ à 95\% (ou 99\%)}}$

2.1. Estimation d'une proportion par IC

Exemple: des biologistes étudient la pollution des océans par les plastiques. Ils veulent déterminer la proportion p de fulmars (espèce d'oiseaux marins) ayant ingéré des matières plastiques.

On appelle X la VA égale à 1 si un oiseau a ingéré du plastique et à 0 sinon. On considère que cette variable suit une loi de Bernoulli B(1, p) où p est le paramètre à estimer (proportion).

Les biologistes examinent un échantillon de 100 cadavres de fulmars trouvés sur le littoral. L'estimateur de p utilisé dans cet échantillon est la fréquence f_n des oiseaux contaminés.

Traitement de l'exemple

Dans l'échantillon aléatoire simple de 100 fulmars, on trouve

$$f_n = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{100}\left(1+0+0+1+1+1+0+....+0+1\right) = 0.85$$

NB: x_i est la $i^{\text{ème}}$ réalisation de X dans l'échantillon, et n = 100

La valeur 0.85 est-elle une bonne approximation de *p*, la proportion inconnue d'oiseaux contaminés dans la population?

Pour répondre à cette question, Il faut en premier lieu trouver la loi de f_n.

Détermination de la loi de f_n

Dans un échantillon aléatoire simple, chaque x_i peut être vu comme la réalisation d'une VA X_i (i = 1, …, n), les X_i étant i.i.d. On rappelle que "i.i.d." signifie:

a. que les VA X_i sont indépendantes deux à deux
b. que chaque VA X_i est de même loi que X, ici *B*(1, *p*)

On sait que la somme de *n* VA de Bernoulli de paramètre *p* est une VA binomiale de paramètres *n* et *p*, *B*(*n*, *p*). Pour *n* > 50, cette loi peut être approchée par une loi normale. On en déduit:

$$\sum_{i=1}^{n} X_i \sim B(n, p) \equiv N\left(np, \sqrt{npq}\right) \qquad \text{avec } q = 1 - p$$

Détermination de la loi de f_n (suite)

On connaît maintenant la loi de la somme des X_i. Comme

$$n.f_n = \sum_{i=1}^{n} X_i \,,$$

on en déduit:

$$nf_n \sim N\left(np, \sqrt{npq}\right) \qquad \text{et donc } f_n \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Il ne reste plus qu'à fixer un niveau de confiance, par exemple 95%, et à trouver les valeurs a et b telles que:

$$\Pr(a \le f_n \le b) = \eta = 0.95$$

Utilisation de la loi *N*(0,1)

Comme $f_n \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$, alors $Z = \dfrac{f_n - p}{\sqrt{\frac{pq}{n}}} \sim \boldsymbol{N}(0, 1)$

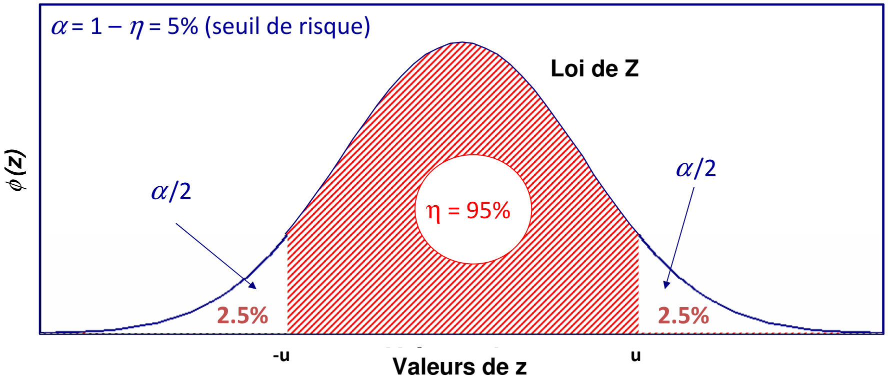
et on peut écrire:

$$\Pr(a \le f_n \le b) = \Pr\left(\frac{a-p}{\sqrt{\frac{pq}{n}}} \le \frac{f_n - p}{\sqrt{\frac{pq}{n}}} \le \frac{b-p}{\sqrt{\frac{pq}{n}}}\right) = 0.95$$

$$\Leftrightarrow \Pr\left(-u \le \frac{f_n - p}{\sqrt{\frac{pq}{n}}} \le u\right) = 0.95 \;\Leftrightarrow \Pr(-u \le Z \le u) = 0.95$$

Détermination de -*u* et *u*

Les valeurs –u et u sont les quantiles de la loi N(0, 1) tels que l'aire comprise sous la courbe entre –u et u soit égale à 0.95:



Lecture de la table de ***N***(0, 1)

On peut trouver -u et u par lecture de la table de la loi ***N***(0,1)

On sait que $\forall Z$, $\Pr(-u \leq Z \leq u) = \Pr(Z \leq u) - \Pr(Z \leq -u)$
$= \Phi(u) - \Phi(-u)$.

où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$. Par symétrie, on a $\Phi(-u)=1-\Phi(u)$, d'où $\Pr(-u \leq Z \leq u) = 2\Phi(u) - 1$

Comme on a choisi $\Pr(-u \leq Z \leq u) = 0.95$, il suffit de lire dans la table la valeur de *u* t.q. $2\Phi(u) - 1 = 0.95$, c-à-d. t.q. $\Phi(u) = 0.975$

On trouve u ≈ 1.96 et donc –u ≈ -1.96 par symétrie

Construction de l'IC

On sait maintenant qu'avec 95% de chances:

$$-1.96 \leq \frac{f_n - p}{\sqrt{\frac{pq}{n}}} \leq 1.96$$

c'est-à-dire $f_n - 1.96\sqrt{\frac{pq}{n}} \leq p \leq f_n + 1.96\sqrt{\frac{pq}{n}}$

On a un encadrement de p dans lequel p et $q = 1 - p$ sont inconnus. On remplace p par son estimateur f_n pour calculer les bornes, ce qui donne finalement l'IC à 95%:

$$f_n - 1.96\sqrt{\frac{f_n(1-f_n)}{n}} \leq p \leq f_n + 1.96\sqrt{\frac{f_n(1-f_n)}{n}}$$

Application numérique

Dans notre exemple, en remplaçant f_n par 0.85 et n par 100, on trouve $0.78 \leq p \leq 0.92$ au niveau de confiance 0.95.

Autrement dit, il y a 95% de chances que la proportion d'oiseaux contaminés par des matières plastiques dans la population soit comprise entre 0.78 et 0.92.

On a 5% de chances de se tromper en donnant cette conclusion.

On remarque que la proportion trouvée dans l'échantillon (0.85) est bien comprise entre 0.78 et 0.92. On peut en conclure que cette valeur est une bonne estimation de p.

IC pour une proportion: formule générale

L'intervalle de confiance pour une proportion p se calcule à partir de la fréquence d'échantillon f_n (où n = taille d'échantillon).

Pour un niveau de confiance η (c'est-à-dire un seuil de risque α) la formule de cet IC est:

$$\boxed{f_n - u\sqrt{\frac{f_n(1-f_n)}{n}} \le p \le f_n + u\sqrt{\frac{f_n(1-f_n)}{n}}}$$

où $u = \Phi^{-1}(\frac{\eta+1}{2}) = \Phi^{-1}(1-\frac{\alpha}{2})$ avec Φ fonction de répartition de ***N***(0,1).

Pour $\eta = 0.95$, on a $u \approx 1.96$ et pour $\eta = 0.99$ on a $u \approx 2.58$

IC pour une proportion: intervalles unilatéraux

Si, au lieu d'encadrer *p*, on cherche une borne (supérieure ou inférieure) à *p*, on construit un intervalle unilatéral. On distinguera:

- les intervalles unilatéraux à gauche, de la forme $0 \le p \le b$, dont la formule générale est:

$$0 \le p \le f_n + u\sqrt{\frac{f_n(1-f_n)}{n}}$$

- les intervalles unilatéraux à droite, de la forme $a \le p \le 1$, dont la formule générale est :

$$f_n - u\sqrt{\frac{f_n(1-f_n)}{n}} \le p \le 1$$

Dans les deux cas, on prendra $u = \Phi^{-1}(\eta) = \Phi^{-1}(1-\alpha)$ avec Φ fonction de répartition de ***N***(0,1).

2.2. Espérance d'une loi *N*(μ,σ)

Dans une population données, soit une VA X de loi ***N***(μ,σ). On cherche à estimer le paramètre inconnu μ par intervalle de confiance.

Dans un échantillon aléatoire simple de taille n tiré de cette population, on prend pour estimateur de μ la moyenne *X* "barre"

Pour encadrer μ, il faut considérer deux cas possibles:

- la variance de la population σ^2 est connue (ce qui n'est en général pas le cas)
- la variance de la population σ^2 est inconnue (ce qui est le cas le plus général)

Cas 1: la variance σ^2 est connue

Quand la variance σ^2 est connue, la formule de l'IC de l'espérance μ pour un niveau de confiance η est :

$$\boxed{\overline{X} - u\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + u\frac{\sigma}{\sqrt{n}}}$$

où σ est l'écart-type de la population et $u = \Phi^{-1}[(\eta+1)/2] = \Phi^{-1}(1-\alpha/2)$. La valeur de u est donc déterminée comme dans le calcul d'un IC pour une proportion.

On peut également chercher un IC unilatéral à gauche ou à droite; la manière de procéder est la même que pour une proportion, excepté que μ n'est pas bornée par 0 ni par 1.

Calcul de la formule de l'IC bilatéral

On prend "X barre" pour estimateur de l'espérance μ. On sait (chapitre 3) que la distribution de "X barre" est approximativement normale, d'espérance μ et de variance σ^2/n.

On a donc $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

Pour un IC bilatéral au niveau de confiance η, on écrit:

$$\Pr\left(-u \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u\right) = \eta \Leftrightarrow \Phi(u)-\Phi(-u)=\eta \Leftrightarrow \Phi(u) = \frac{\eta+1}{2}$$

On a donc, avec un niveau de confiance η:

$$-u \le \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u \quad \text{c'est-à-dire} \quad \boxed{\overline{X} - u\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + u\frac{\sigma}{\sqrt{n}}}$$

Calcul des formules des IC unilatéraux

Pour un IC unilatéral à droite au niveau de confiance η, on écrit:

$$\Pr\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u\right) = \eta \Leftrightarrow \Phi(u) = \eta$$

On a donc, avec un niveau de confiance η : $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u$

c'est-à-dire $\boxed{\overline{X} - u\dfrac{\sigma}{\sqrt{n}} \le \mu}$

Pour un IC unilatéral à gauche au niveau de confiance η, on écrit:

$$\Pr\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \ge -u\right) = \eta \Leftrightarrow \Pr\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le u\right) = \eta \Leftrightarrow \Phi(u) = \eta$$

On a donc, avec un niveau de confiance η : $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \ge -u$

c'est-à-dire $\boxed{\mu \le \overline{X} + u\dfrac{\sigma}{\sqrt{n}}}$

Cas 2: la variance σ^2 est inconnue

Quand la variance σ^2 est inconnue, la formule de l'IC de l'espérance μ pour un niveau de confiance η est :

$$\boxed{\overline{X} - t_{n-1}\frac{S'}{\sqrt{n-1}} \le \mu \le \overline{X} + t_{n-1}\frac{S'}{\sqrt{n-1}}}$$

où S' est l'estimateur du MV de σ et $t_{n\text{-}1}$ le quantile d'ordre (η+1)/2 de la loi de Student à n-1 d.l., *St*(n-1) : $t_{n\text{-}1} = F^{-1}[(\eta+1)/2] = F^{-1}[1-\alpha/2]$, avec *F* fonction de répartition de *St*(n-1)

La formule de S' est $S' = \hat{\sigma} = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2}$, qui peut s'exprimer en fonction de l'écart-type empirique S (cf. Chapitre 1).

Ecriture avec S au lieu de S'

L'estimateur S' peut se réécrire:

$$S' = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2} = \sqrt{\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2}$$

$$= \sqrt{\frac{n-1}{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2} = \sqrt{\frac{n-1}{n}}S$$

Ce qui donne: $\dfrac{S'}{\sqrt{n-1}} = \dfrac{S}{\sqrt{n}}$. La formule de l'IC de μ devient alors:

$$\boxed{\overline{X} - t_{n-1}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1}\frac{S}{\sqrt{n}}}$$

Valeurs usuelles de t_{n-1}

On peut trouver les valeurs de t_{n-1} à l'aide des tables de la loi de Student ou du logiciel Excel (function "LOI.STUDENT.INVERSE()" qui utilise le seuil de risque α et non le niveau de confiance η).

Rappel: l'allure de la distribution de Student est similaire à celle de la loi normale $\boldsymbol{N}(0, 1)$. Quand $n \to +\infty$ (en pratique, pour n>100), la loi de Student tend vers la loi normale centrée réduite $\boldsymbol{N}(0, 1)$.

On aura donc pour n "grand": $t_{n-1} = 1.96$ si $\eta=0.95$
$t_{n-1} = 2.48$ si $\eta=0.99$

Intervalles unilatéraux

La formule de l'intervalle unilatéral à gauche de μ (σ² inconnue) est:

$$\mu \le \overline{X} + t_{n-1}\frac{S}{\sqrt{n}}$$

La formule de l'intervalle unilatéral à droite μ (σ² inconnue) est:

$$\overline{X} - t_{n-1}\frac{S}{\sqrt{n}} \le \mu$$

Dans les deux cas, on prendra $t_{n-1} = F^{-1}(\eta) = F^{-1}(1-\alpha)$, avec F fonction de répartition de $St(n-1)$.

Calcul de la formule de l'IC bilatéral

On prend la moyenne "X barre" pour estimateur de l'espérance μ, et S', estimateur du MV de σ, pour l'écart-type. Le point de départ du calcul est le **Théorème de Fisher**, selon lequel:

$$\frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \sim St(n-1)$$

Pour un IC bilatéral au niveau de confiance η, on écrit:

$$\Pr\left(-t \le \frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \le t\right) = \eta \Leftrightarrow F(t) - F(-t) = \eta \Leftrightarrow F(t) = \frac{\eta+1}{2}$$

où F est la fonction de répartition de $St(n-1)$. On a $t = t_{n-1} = F^{-1}[(\eta+1)/2]$.
Au niveau de confiance η, on peut donc écrire:

$$-t_{n-1} \le \frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \le t_{n-1} \text{, soit : } \boxed{\overline{X} - t_{n-1}\frac{S'}{\sqrt{n-1}} \le \mu \le \overline{X} + t_{n-1}\frac{S'}{\sqrt{n-1}}}$$

Calcul des formules des IC unilatéraux

Pour un IC unilatéral à droite au niveau de confiance η, on écrit:

$$\Pr\left(\frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \le t_{n-1}\right) = \eta \Leftrightarrow F(t_{n-1}) = \eta \Leftrightarrow t_{n-1} = F^{-1}(\eta)$$

On a donc, avec un niveau de confiance η : $\frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \le t_{n-1}$

c'est-à-dire $\boxed{\overline{X} - t_{n-1}\frac{S'}{\sqrt{n-1}} \le \mu}$

Pour un IC unilatéral à gauche au niveau de confiance η, on écrit:

$$\Pr\left(\frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \ge -t_{n-1}\right) = \eta \Leftrightarrow \Pr\left(\frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \le t_{n-1}\right) = \eta \Leftrightarrow F(t_{n-1}) = \eta$$

On a donc, avec un niveau de confiance η : $\frac{\overline{X}-\mu}{S'/\sqrt{n-1}} \ge -t_{n-1}$

c'est-à-dire $\boxed{\mu \le \overline{X} + t_{n-1}\frac{S'}{\sqrt{n-1}}}$

Illustration

Les biologistes de l'exemple 2.1 veulent maintenant déterminer la quantité moyenne de plastique ingérée par les fulmars.

On appelle X la VA continue représentant la quantité de plastique ingérée (en g). Dans la population, cette variable suit une loi normale N(μ, σ) de paramètres inconnus. Les biologistes veulent estimer μ au moyen d'un IC.

L'estimateur de μ est la moyenne "X barre". Dans l'échantillon de 100 fulmars, les biologistes trouvent une moyenne de 0.6 g et un écart-type empirique de 0.3 g de plastique.

Application numérique

On se donne un niveau de confiance de 95%. A ce niveau de confiance et pour n-1 = 99 d.l., on trouve t_{n-1}=1.98

On applique alors la formule de l'IC quand la variance est inconnue

$$\overline{X} - t_{n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1}\frac{S}{\sqrt{n}}$$

aux valeurs trouvées dans l'échantillon de 100 fulmars. Il vient:

$$0{,}6 - 1{,}98\frac{0{,}3}{\sqrt{100}} \leq \mu \leq 0{,}6 + 1{,}98\frac{0{,}3}{\sqrt{100}} \quad \Leftrightarrow \quad \boxed{0{,}54 \leq \mu \leq 0{,}66}$$

"La quantité moyenne de plastique ingérée par les oiseaux est comprise entre 0.54 g et 0.66 g avec une probabilité de 95%"

2.3. Variance d'une loi *N*(μ,σ)

Dans une population données, soit une VA X de loi ***N***(μ,σ). On cherche un IC pour estimer le paramètre inconnu σ^2.

Dans un échantillon aléatoire simple de taille n tiré de cette population, on prend pour estimateur de σ^2 l'estimateur du MV:

- $S'^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ si l'espérance μ est connue, ce qui n'est en général pas le cas.
- $S'^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$ si l'espérance μ est inconnue, ce qui est le cas le plus général (on remplace μ par son estimateur "X barre").

Cas 1: l'espérance μ est connue

On utilise l'estimateur de σ^2 dont la formule fait intervenir μ. La formule de l'IC de la variance σ^2 pour un niveau de confiance η est :

$$\boxed{\frac{n\hat{\sigma}^2}{k_1} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{k_2}}$$

où k_1 et k_2 sont des quantiles de la loi du Khi-2 à n d.l. $\chi^2(n)$. On prend en général $k_1 = F^{-1}(1-\alpha/2)$ et $k_2 = F^{-1}(\alpha/2)$, avec F fonction de répartition de $\chi^2(n)$ et $\alpha = 1 - \eta$.

Pour un IC unilatéral à gauche on prendra $k_2 = F^{-1}(\alpha)$ et pour un IC unilatéral à droite on prendra $k_1 = F^{-1}(1-\alpha)$. On aura:

$$0 \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{k_2} \quad \text{et} \quad \frac{n\hat{\sigma}^2}{k_1} \leq \sigma^2$$

Calcul de la formule de l'IC

Comme μ est connue, on part de la formule de l'estimateur :

$$S'^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

On utilise la propriété (rencontrée lors du calcul de l'espérance de l'estimateur) : si X_1, …, X_n est une suite de VA i.i.d. de loi ***N***(0,1), alors

$$\sum_{i=1}^{n} X_i^{\,2} \sim \chi^2(n)$$

On remarque que $nS'^2 = \sum_{i=1}^{n}(x_i - \mu)^2$ et donc $\frac{nS'^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$

Comme $\forall i, \frac{x_i - \mu}{\sigma} \sim N(0, 1)$, on en déduit que $\frac{nS'^2}{\sigma^2} \sim \chi^2(n)$

Calcul de la formule de l'IC bilatéral

Pour un IC bilatéral au niveau de confiance η, on pose:

$$\Pr\left(k_2 \le \frac{nS'^2}{\sigma^2} \le k_1\right) = \eta = 1-\alpha \Leftrightarrow \Pr\left(\frac{nS'^2}{\sigma^2} \le k_1\right) - \Pr\left(\frac{nS'^2}{\sigma^2} \le k_2\right) = \eta$$

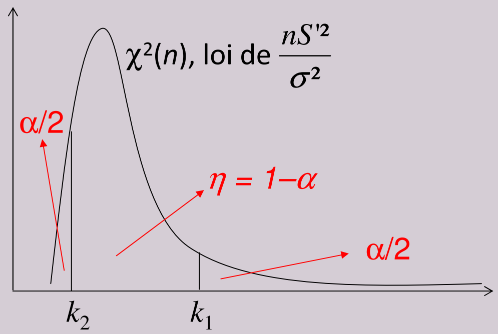
$$\Leftrightarrow F(k_1) - F(k_2) = (1-\alpha/2) - \alpha/2 = \eta$$

On a donc, au niveau de confiance η:

$$k_2 \le \frac{nS'^2}{\sigma^2} \le k_1$$

c'est-à-dire

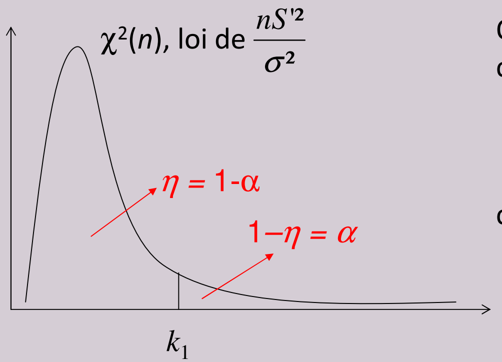
$$\boxed{\frac{nS'^2}{k_1} \le \sigma^2 \le \frac{nS'^2}{k_2}}$$



Calcul de la formule de l'IC unilatéral à droite

Pour un IC unilatéral à droite au niveau de confiance η, on pose:

$$\Pr\left(\frac{nS'^2}{\sigma^2} \le k_1\right) = F(k_1) = \eta = 1-\alpha$$



On a donc, au niveau de confiance η:

$$\frac{nS'^2}{\sigma^2} \le k_1$$

c'est-à-dire

$$\boxed{\frac{nS'^2}{k_1} \le \sigma^2}$$

Calcul de la formule de l'IC unilatéral à gauche

Pour un IC unilatéral à gauche au niveau de confiance η, on pose:

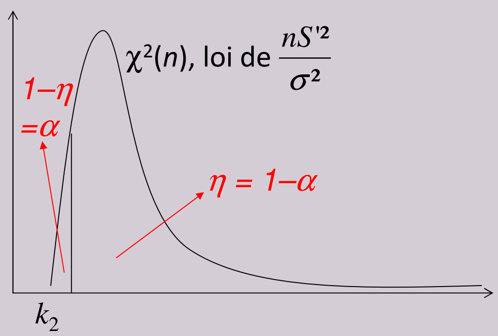
$$\Pr\left(\frac{nS'^2}{\sigma^2} \ge k_2\right) = \eta = 1-\alpha \Leftrightarrow \Pr\left(\frac{nS'^2}{\sigma^2} \le k_2\right) = 1-\eta = \alpha$$

On a donc, au niveau de confiance η:

$$\frac{nS'^2}{\sigma^2} \ge k_2$$

c'est-à-dire

$$\boxed{\sigma^2 \le \frac{nS'^2}{k_2}}$$



Cas 2: l'espérance μ est inconnue

On utilise l'estimateur de σ^2 dont la formule fait intervenir "*X* barre". La formule de l'IC de σ^2 pour un niveau de confiance η est :

$$\boxed{\frac{n\hat{\sigma}^2}{k_1} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{k_2} \Leftrightarrow \frac{(n-1)S^2}{k_1} \leq \sigma^2 \leq \frac{(n-1)S^2}{k_2}}$$

avec $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$ variance empirique de l'échantillon.

Cette fois, k_1 et k_2 sont des quantiles de la loi $\chi^2(n\text{-}1)$, $k_1 = F^{-1}(1-\alpha/2)$ et $k_2 = F^{-1}(\alpha/2)$, avec F fonction de répartition de $\chi^2(n\text{-}1)$.

Comme dans le cas 1, on prend $k_2 = F^{-1}(\alpha)$ pour un IC unilatéral à gauche et $k_1 = F^{-1}(1-\alpha)$ pour un IC unilatéral à droite, en adaptant la formule ci-dessus de la manière habituelle.

Illustration

Les biologistes de l'exemple 2.1 veulent maintenant estimer la variance de la quantité de plastique ingérée par les fulmars.

La VA continue représentant la quantité de plastique ingérée (en g) est toujours X, de loi $N(\mu, \sigma)$ dans la population. Les paramètres μ et σ sont inconnus. Les biologistes veulent estimer σ^2 au moyen d'un IC.

L'espérance est inconnue, l'estimateur de σ^2 utilise donc la moyenne empirique "X barre". Dans l'échantillon de 100 fulmars, les biologistes trouvent une moyenne de 0.6 g et un écart-type empirique de 0.3 g de plastique (ce qui donne une variance empirique de 0.09).

Application numérique

On se donne un niveau de confiance de 95%. A ce niveau de confiance et pour n-1 = 99 d.l., on trouve k_1 = 128,42 et k_2 = 73,37.

On applique alors la formule de l'IC

$$\frac{n\hat{\sigma}^2}{k_1} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{k_2} \Leftrightarrow \frac{(n-1)S^2}{k_1} \leq \sigma^2 \leq \frac{(n-1)S^2}{k_2}$$

aux valeurs trouvées dans l'échantillon de 100 fulmars. Il vient:

$$\frac{99\times 0{,}09}{128{,}42} \leq \sigma^2 \leq \frac{99\times 0{,}09}{73{,}37} \quad \Leftrightarrow \quad \boxed{0{,}07 \leq \sigma^2 \leq 0{,}12}$$

"La variance de la quantité de plastique ingérée par les oiseaux est comprise entre 0,07 et 0,12 avec une probabilité de 95%"