Chap 1 - Estimateurs et Intervalles de confiance

Jordy Palafox

September 28, 2023

Statistiques Inférentielles
CY Tech - Ing 2 GSI 2023-2024

Motivations

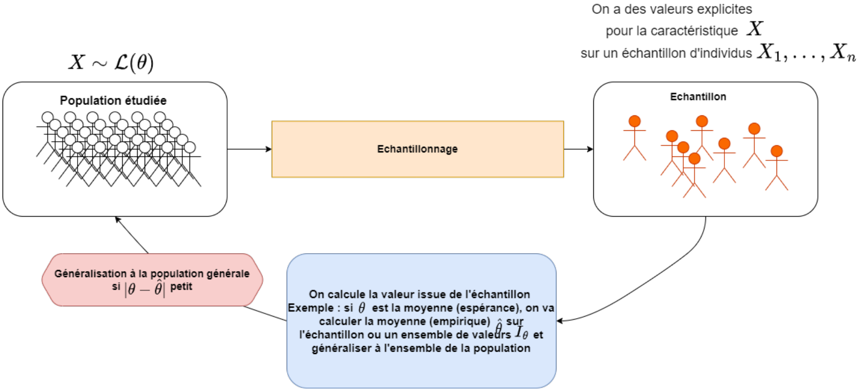
- On s'intéresse à la caractéristique d'une population modélisée par une variable X donc on connait la loi (binomiale, exponentielle, gaussienne etc...)
- Cette loi dépend d'un paramètre :
  1. p pour la loi binomiale,
  2. λ pour la loi de Poisson,
  3. θ pour la loi exponentielle,
  4. μ et σ pour la loi normale,
  5. ...

Problème

Ce paramètre est **inconnu**, on va devoir l'évaluer, l'approcher, l'estimer à partir d'un échantillon de la population :

- soit par une valeur estimée (approché) $\hat{\theta}$, on parle d'**estimation ponctuelle**,
- soit on se donne un intervalle de valeurs dont on est "sûr" avec un certain niveau de confiance qu'il contient θ, on parle d'**intervalle de confiance** .

En résumé



On a des valeurs explicites
pour la caractéristique X
sur un échantillon d'individus $X_1, \ldots, X_n$
$X \sim \mathcal{L}(\theta)$
Population étudiée
Echantillonnage
Echantillon
Généralisation à la population générale
si $|\theta - \hat{\theta}|$ petit
On calcule la valeur issue de l'échantillon
Exemple : si θ est la moyenne (espérance), on va
calculer la moyenne (empirique) $\hat{\theta}$ sur
l'échantillon ou un ensemble de valeurs I_θ et
généraliser à l'ensemble de la population

Exemple

On s'intéresse au temps d'attente entre deux atterrissages sur un aéroport, modélisé par une variable aléatoire X qui suit une loi exponentielle : $X \sim \mathcal{E}(\theta)$.
On dispose des 9 observations suivantes :

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.4	1	0.5	0.3	0.1	0.2	1.4	0.2	0.5

Comment estimer θ (qui est inconnu !) ?

Définition

Soit X une variable aléatoire et $X_1, \dots, X_n$ une échantillon i.i.d. de X (i.e $\mathcal{L}_{X_i} \sim \mathcal{L}_X$).

- Une **statistique** T est une v.a. fonction de $(X_1, ..., X_n)$ i.e $T = T(X_1, \dots, X_n)$ dont on observe une réalisation $t = T(x_1, \dots, x_n)$.
- Une statistique utilisée pour estimer un paramètre θ est appelée **estimateur**, notée souvent $\hat{\theta}_n$.

Application à notre exemple

Exemple suite

Une loi exponentielle $\mathcal{E}(\theta)$ admet pour espérance et variance (théorique) :

$$\mathbb{E}(X) = \frac{1}{\theta} \; et \; \mathbb{V}ar(X) = \frac{1}{\theta^2}$$

On peut donc estimer θ de deux manières :

- soit par l'**espérance empirique** (estimateur de l'espérance) $\bar{X}_n = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, on a donc :

  $$\frac{1}{9}(0.4 + 1 + 0.5 + 0.3 + 0.1 + 0.2 + 1.4 + 0.2 + 0.5) = 0.51111$$

- soit par la **variance empirique** (estimateur de la variance) $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ ce qui nous donne (avec R) 0.181111.
  Pourtant à la main on trouve 0.1609877 !

Application à notre exemple

Exemple suite

- On trouve $\bar{x} = 0.511111$, on trouve $\hat{\theta} = \frac{1}{\bar{x}} \simeq 1.96$
- Avec la variance empirique, on trouve $\hat{\theta} = \frac{1}{\bar{x}^2} \simeq 2.35$

Questions

- Pourquoi la valeur de la variance est différente avec R du calcul à la main ?
- Quelle est la valeur de $\hat{\theta}$ à retenir ?

On va évaluer la qualité des estimateurs !

Qualité d'un estimateur

Soit $\hat{\theta}$ un estimateur du paramètre θ, notre objectif est de reduire l'erreur d'estimateur $|\hat{\theta} - \theta|$.

Définition

Le **biais** d'un estimateur $\hat{\theta}$ du paramètre θ est défini par :

$$b(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta) = \mathbb{E}(\hat{\theta}) - \theta$$

Un estimateur est dit **sans biais** si $b(\hat{\theta}) = 0$, l'erreur moyenne est nulle i.e $\mathbb{E}(\hat{\theta}) = \theta$.

Suite de l'exemple

L'estimateur de la variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ est biaisé, vérifiez-le!

On préfèrera la **variance empirique non biaisée** $v^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}s^2$, qui est sur R (cf data explo!).

Qualité d'un estimateur

Définition

Le **risque quadratique** d'un estimateur $\hat{\theta}$ du paramètre θ est défini par :

$$R(\hat{\theta}) = \mathbb{E}\left((\hat{\theta} - \theta)^2\right).$$

On l'appelle aussi **erreur quadratique moyenne** et on trouvera aussi la notation $EQM(\hat{\theta})$.

Décomposition biais-variance

$$R(\hat{\theta}) = \mathbb{V}ar(\hat{\theta}) + b(\hat{\theta})^2.$$

Notion de convergence : a-t-on $\lim\limits_{n \to +\infty} \hat{\theta}_n = \theta$?

Définition

Un estimateur est dit **convergent** si $\lim\limits_{n \to +\infty} R(\hat{\theta}_n) = 0$.

Il existe d'autres types de convergences : en probabilité ou presque-sûre qui découlent des lois des grands nombres, **gardez-les à l'esprit**!

Comment choisir entre deux estimateurs ?

1. On garde celui sans biais,
2. Puis on garde celui qui a la plus petite variance !

Exemple de comparaison d'estimateurs

Application

On considère $X \sim \mathcal{U}_{[0,\theta]}$ avec $\theta > 0$ un paramètre inconnu et un échantillon de taille n de X : X_1, ..., X_n. Soit les trois estimateurs:

$$T_n = 2\bar{X}_n, \quad T'_n = max(X_1, ..., X_n) \;\; et \;\; T''_n = \frac{n+1}{n} T'_n.$$

On rappelle que $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. Rappeler l'espérance de X.
2. Donner la fonction de répartition de T', sa densité, l'espérance et la variance.
3. Calculer le biais des estimateurs, que se passe-t-il pour n grand ?
4. Calculer le risque quadratique .
5. Conclure sur le meilleur estimateur de θ.

Estimateur d'une proportion

Dans le cas d'une **proportion**, on va distinguer une catégorie spécifique parmis la population étudiée. Soit on vérifie le critère soit non : "abimé ou non", "neuf ou non", "jeune ou non" ... C'est donc modélisé par une épreuve de Bernouilli : $X \sim \mathcal{B}(p)$. Ici p est le paramètre à estimer.

L'estimateur usuel d'une proportion est la **fréquence empirique** :

$$F_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Biais** : $b(p) = \mathbb{E}(F_n) - p = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) - p = \frac{1}{n}np - p = 0.$
  C'est un **estimateur sans biais** !
- **Risque empirique** : $R(F_n) = \mathbb{V}ar(F_n) = \frac{\mathbb{V}ar(X)}{n} = \frac{p(1-p)}{n}$ qui tend vers 0 en $n\infty$, donc **convergent**.

Estimateur d'une moyenne

On a déjà fait le travail ! Résumons :
Soit X une variable aléatoire de variance σ^2 avec $\mathbb{E}(X) = \mu$ le paramètre à estimer.

L'estimateur de l'espérance est la **moyenne empirique** :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Biais** : $b(\mu) = \mathbb{E}(\bar{X}_n) - \mu = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) - \mu = 0$. C'est un **estimateur sans biais** !
- **Risque quadratique** : $R(\bar{X}_n) = \mathbb{V}ar(\bar{X}_n) = \frac{\sigma^2}{n}$ qui est aussi **convergent**.

Estimateur d'une variance

Soit X une variable aléatoire telle que $\mathbb{V}ar(X) = \sigma^2$ est le paramètre à estimer.

- L'estimateur usuel de la variance est la **variance empirique**

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - (\bar{X}_n)^2.$$

qui est **biaisé** : $b(S^2) = \frac{-1}{n+1}\sigma^2$.

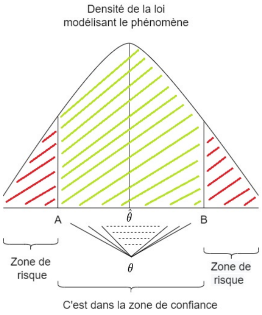
- On utilisera souvent la **variance empirique corrigée**, qui est **sans biais** :

$$S^{*2} = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

Les deux estimateurs **convergent presque sûrement** mais en moyenne quadratique seulement **dans le cas gaussien**.

Motivation

On vient de voir comment déterminer une valeur explicite et approchée $\hat{\theta}$ du paramètre θ d'une loi permettant de modéliser un phénomène.
Plutôt que de déterminer une seule valeur $\hat{\theta}$ on va chercher un ensemble de valeur (**intervalle de confiance**) dont on est presque sûr (**niveau de confiance**) qu'il contient la valeur théorique !



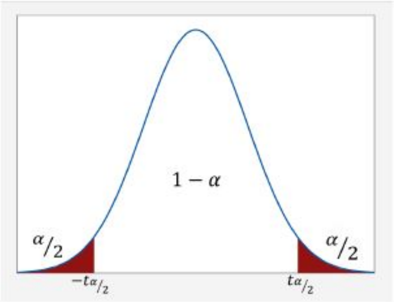
Intérêt : l'intervalle de confiance permet de quantifier la probabilité d'erreur dans le choix de l'estimateur.

Plus concrètement ...

1. On a un paramètre (théorique) θ d'une loi à estimer.
2. On va se fixer un **risque** noté α qui nous dit : "mon intervalle a 95% de chance de contenir le paramètre".
3. On va chercher donc les extrémités A et B d'un intervalle tel que $\mathbb{P}(\theta \in [A, B]) = 1 - \alpha$.

Vocabulaire :

- α désigne le **risque**,
- $1 - \alpha$ désigne le **niveau de confiance**.



Procédure générale

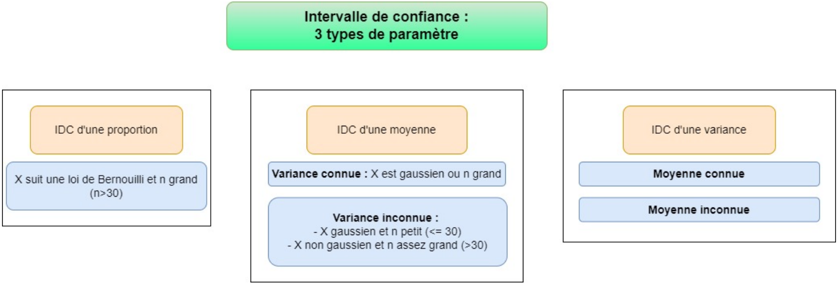
En pratique

- On choisit un estimateur $\hat{\theta}$ de θ et on va construire un intervalle **symétrique** en $\hat{\theta}$ i.e $A = \hat{\theta} - \ell$ et $B = \hat{\theta} + \ell$.
- On se fixe un niveau de risque α (en général $1,\ 2.5,\ 5\%$)
- On a donc :

$$\begin{aligned}\mathbb{P}(\hat{\theta} - \ell \leq \theta \leq \hat{\theta} + \ell) &= \mathbb{P}(\theta - \ell \leq \hat{\theta} \leq \theta + \ell) \\ &= 1 - \alpha.\end{aligned}$$

- On détermine alors ℓ grâce à la table de la loi de l'estimateur $\hat{\theta}$.

Les cas rencontrés



IDC d'une proportion p

Supposons que $X \sim \mathcal{B}(p)$ et que l'échantillon est suffisament grand i.e $n > 30$. On va utiliser l'estimateur empirique de la fréquence $F_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ pour approcher p :

Procédure

1. On commence par remarquer que comme $n > 30$, on peut utiliser le TCL : $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} = \frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{n} \sim \mathcal{N}(0,1)$.
2. On se fixe un risque α tel que $\mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) = 1 - \alpha$
3. Il suffit de lire le quantile t_α associé à la probabilité $1 - \alpha$ dans la table de la loi $\mathcal{N}(0,1)$.
4. Bien remarquer que on a un IDC sur Z, il faut donc isoler p :

$$\begin{aligned}
1 - \alpha &= \mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) \\
&= \mathbb{P}\left(-t_\alpha \leq \frac{F_n - p}{\sqrt{p(1-p)}}\sqrt{p} \leq t_\alpha\right)
\end{aligned}$$

IDC pour une proportion

Procédure

4. $$= \mathbb{P}\left(F_n - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq F_n + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

5. On obtient l'IDC :

$$[A, B] = \left[F_n - t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, F_n + t_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

6. On remarque que l'IDC de p dépend de p lui-même, on va donc estimer F_n sur l'échantillon et on remplace par la valeur estimée f_n dans l'IDC :

$$[a, b] = \left[f_n - t_\alpha \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}, f_n + t_\alpha \frac{\sqrt{f_n(1-f_n)}}{\sqrt{n}}\right]$$

Enoncé

On dispose d'un échantillon de 503 pommes mais 47 sont abîmées. Déterminer un intervalle de confiance de la porportion de pommes aîmées sur l'ensemble du verger.

Correction

1. On considère $X \sim \mathcal{B}(p)$ où X modélise : la pomme est-elle abimée ou non. De plus l'échantillon est assez grand ($503 > 30$) donc on est dans les hypothèses de la procédure.
2. On va commencer par calculer le proportion de pommes abîmées observée $f_n = \frac{47}{503} \simeq 0.09343936 = 9.343936\%$.
3. Fixons nous un risque $\alpha = 5\%$. Ainsi le quantile $t_\alpha = 1.96$ d'après la table de la loi $\mathcal{N}(0, 1)$.
4. Alors l'IDC de p est :

$$[a, b] = [6.800412\%, \ 11.88746\%]$$

Conclusion

La proportion de pommes abîmées dans le verger est donc est 6.8% et 11.88% avec un risque d'erreur de 5% (ou une confiance de 95%)

IDC pour une moyenne, variance σ^2 connue

$\boxed{X \text{ suit une loi gaussienne } \mathcal{N}(\mu, \sigma^2) \text{ ou } n \text{ grand } (n > 30)}$

On utilise l'estimateur $\bar{X}_n$ estimateur de μ et σ^2 est connue (on a une valeur explicite).

Procédure

1. On considère $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ ou de manière équivalente $\dfrac{\overline{X_n} - \mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$ (Démontrer le)
2. Il suffit de lire sur la table de $\mathcal{N}(0,1)$ le quantile t_α tel que $\mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) = 1 - \alpha$.
3. On a ainsi :

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) \\ &= \mathbb{P}\left(-t_\alpha \leq \frac{\overline{X_n} - \mu}{\sigma}\sqrt{n} \leq t_\alpha\right) \end{aligned}$$

IDC pour une moyenne, variance σ^2*connue*

Procédure

3. Ainsi :

$$1 - \alpha = \mathbb{P}\left(\overline{X_n} - t_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu \leq + t_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

4. On obtient finalement l'IDC :

$$[a, b] = \left[\bar{x}_n - t_\alpha \frac{\sigma}{\sqrt{n}}, \bar{x}_n + t_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Exercice d'application

Exercice

Une sociète fabrique des câbles d'acier dont on souhaite étudier la résistance moyenne de rupture. Au vue de l'utilisation des câbles, la résistance doit avoir un écart-type $\sigma = 25$kg.
On dispose d'un échantillon de 100 câbles sur lequel on observe une résistance moyenne $\bar{x}$ de 2630 kg.
Déterminons un intervalle de confiance de la résistance moyenne de rupture μ avec un risque 2.5%.

Correction

1. L'échantillon est assez grand ($n > 30$). On considère donc $\overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ donc $\frac{\overline{X_n}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$.
2. On cherche t_α tel que :

$$\begin{aligned} 0.975 &= \mathbb{P}(-t_\alpha \leq Z \leq t_\alpha) \\ &= F_Z(t_\alpha) - F_Z(-t_\alpha) \\ &= 2F_Z(t_\alpha) - 1. \end{aligned}$$

D'où $F_Z(t_\alpha) = 0.9875$ et par la table $\mathcal{N}(0,1)$ on a $t_\alpha = 2.241403$.
On peut maintenant déterminer l'IDC.

IDC pour une moyenne, variance σ^2 inconnue, n petit

$\boxed{X \text{ suit une loi gaussienne } \mathcal{N}(\mu, \sigma^2) \text{ et } n \text{ petit } (n \leq 30)}$

On a toujours $\frac{1}{n}\sum_{i=1}^{n} X_i$ estimateur de $\mu = \mathbb{E}(X)$ et $\sigma^2 = \mathbb{V}ar(X)$.

Procédure (c'est toujours la même chose)

1. On utilise S^{*2} (la variance empirique non biaisée) au lieu de σ^2 (inconnue !),
2. On pose $Y = \dfrac{\overline{X_n} - \mu}{S^*}\sqrt{n}$ telle que $Y \sim \mathcal{T}_{n-1}$ loi de Student à $n-1$ degré de liberté.
3. La table de $\mathcal{T}_{n-1}$ donne t_α tel que $\mathbb{P}(-t_\alpha \leq Y \leq t_\alpha) = 1 - \alpha$,
4. En transformant l'expression de sorte à encadrer μ on obtient l'IDC :

$$[a, b] = \left[\bar{x}_n - t\frac{s^*}{\sqrt{n}},\ \bar{x}_n + t\frac{x^*}{\sqrt{n}}\right]$$

IDC pour une moyenne, variance σ^2 inconnue, n grand

$\boxed{X \text{ est non gaussien et } n \text{ assez grand } (n > 30)}$

On a encore et toujours $\frac{1}{n}\sum_{i=1}^{n} X_i$ estimateur de $\mu = \mathbb{E}(X)$ et $\sigma^2 = \mathbb{V}ar(X)$.

Procédure (c'est toujours la même chose)

1. On utilise S^{*2} (la variance empirique non biaisée) au lieu de σ^2 (inconnue !),
2. On pose $Y = \frac{\overline{X_n} - \mu}{S^*}\sqrt{n}$ telle que $Y \sim \mathcal{N}(0,1)$.
3. La table de $\mathcal{N}(0,1)$ donne t_α tel que $\mathbb{P}(-t_\alpha \leq Y \leq t_\alpha) = 1 - \alpha$,
4. En transformant l'expression de sorte à encadrer μ on obtient l'IDC :
$$[a, b] = \left[\bar{x}_n - t\frac{s^*}{\sqrt{n}},\ \bar{x}_n + t\frac{x^*}{\sqrt{n}}\right]$$

IDC pour une variance, μ connue ou inconnue

$$\boxed{X \text{ suit une loi } \mathcal{N}(\mu, \sigma^2)}$$

Espérance μ **connue**

1. On utilise l'estimateur $\mathbb{V}ar_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$.
2. On a $n\frac{\mathbb{V}ar_1}{\sigma^2} \sim \chi^2_n$.

Espérance μ **inconnue**

1. On utilise l'estimateur $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ ou $S^{*2} = \frac{n}{n-1}S^2$
2. On a $n\frac{S^2}{\sigma^2} = (n-1)\frac{S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$.

Dans les deux cas, c'est la lecture de la table bien choisie qui donne les valeurs de t_α et donc de définir l'IDC.

Exercice

On s'intéresse à la taille des étudiants de CY Tech, sur le campus palois. Mais on ne connait ni l'espérance théorique ni la variance, seulement que la taille X suit une loi gaussienne !
Un échantillon de 25 personnes parmis les ING2 GSI donne les mesures suivantes :

$$\bar{x} = 177.1cm \ et \ s^* = 3.8$$

Donner un IDC à 95% de confiance de la variance σ^2.

Correction

1. D'après ce qu'il précède, $(n-1)\frac{S^{*2}}{\sigma^2} \sim \chi^2_{n-1}$ donc $Y = 24\frac{S^{*2}}{\sigma^2} \sim \chi^2_{24}$.
2. D'après la table de la loi χ^2_{24} on a $t_1 = 12.4$ tel que $0.025 = \mathbb{P}(Y \leq t_1)$,
3. La même table donne $t_2 = 39.36$ tel que $0.975 = \mathbb{P}(Y \leq t_2)$.
4. On en déduit $0.95 = \mathbb{P}(12.4 \leq Y \leq 39.36)$.
5. Après quelques étapes on obtient finalement $0.95 = \mathbb{P}(8.8 \leq \sigma^2 \leq 27.9)$.
   Autrement dit l'IDC est $[8.8,\ 27.9]$.

D'où viennent toutes ces lois et les choix de n ou $n-1$?
Ces choix de lois découlent du **théorème de Cochran**

Enoncé du théorème de Cochran et conséquences

Théorème de Cochran

Soit $X = (X_1, ..., \; X_n)^t$ un vecteur gaussien centré réduit. POur F un sous-espace vectoriel de $\mathbb{R}^n$ de dimension p, on note P_F (resp $P_{F^\perp}$) la projection orthogonale sur F (resp $F^\perp$).
Alors les vecteurs $P_F X$ et $P_{F^\perp} X$ sont gaussiens, indépendants, de lois :

$$P_F X \sim \mathcal{N}(0, P_F) \; et \; P_{F^\perp} X \sim \mathcal{N}(0, P_{F^\perp}).$$

De plus, les variables $\|P_F X\|^2$ et $\|P_{F^\perp} X\|^2$ sont gaussiens, indépendants, de lois :

$$\|P_F X\|^2 \sim \chi^2_p \; et \; \|P_{F^\perp} X\| \sim \chi^2_{n-p}.$$

Corollaire de Cochran

On a : $\overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, $(n-1)\frac{S^2}{\sigma 2} \sim \chi^2_{n-1}$ et $\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} \sim \mathcal{T}_{n-1}$.

Commentaires

1. Le corollaire de Cochran se démontre en utilisant des manipulations proches de celles réalisées en data explo ! Voir `https://perso.univ-rennes1.fr/jean-christophe.breton/agreg/AGREG/COURS/cochran.pdf`
2. Cela explique à la fois le choix des lois et les degrés de liberté.
3. Non, ce n'est pas à connaître, c'est pour votre culture.

A retenir

Si on remplace une valeur théorique par sa valeur estimée sur l'échantillon, on perd un degré de liberté par valeur.