ESTIMATION

Donner une valeur approchée aux paramètres d'une population (μ, σ^2, …)
à l'aide d'un échantillon de cette population

Hypothèse de l'échantillonnage : Les observations $x_1, \dots, x_n$ sont considérées comme des réalisations de $X_1, \dots, X_n$ v.a.i.i.d.

Un estimateur T_n est une variable aléatoire fonction de $X_1, \dots, X_n$
$T_n = f(X_1, \dots, X_n)$

Exemples : moyenne, variance empirique,…

Pour un même paramètre, il peut y avoir plusieurs estimateurs possibles. Lequel choisir?

→ **Qualité d'un estimateur**

ESTIMATION

Exemple

Soit X le temps d'attente entre deux atterrissages sur un aéroport. On peut supposer que X suit une loi exponentielle car temps d'attente. Reste à estimer le paramètre θ d'après les observations

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

On sait $E(X)=1/\theta$ et $var(X)=1/\theta^2$. Donc deux possibilités pour estimer θ

Espérance $\Rightarrow$ $\theta=1.96$

Variance $\Rightarrow$ $\theta=2.35$

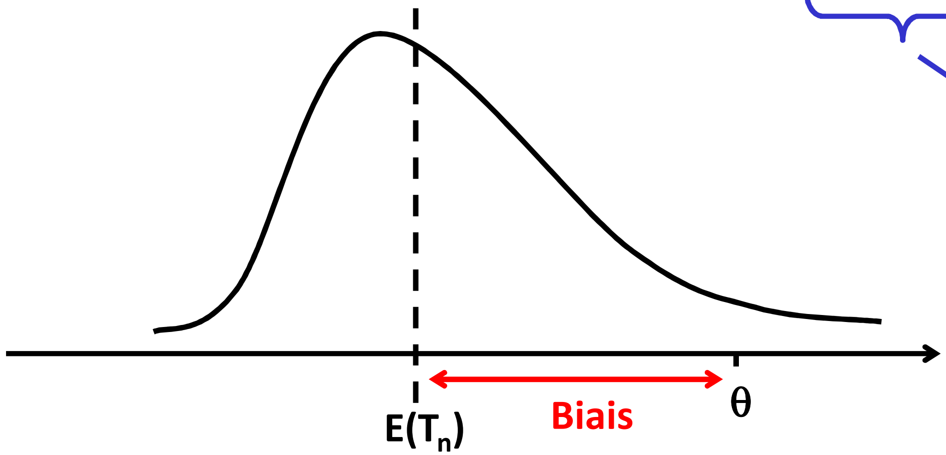
?

Sert ensuite à mieux gérer le temps des contrôleurs

QUALITE D'UN ESTIMATEUR

Soit θ le paramètre à estimer et T_n un estimateur de θ. On étudie l'écart $T_n - \theta$

$$T_n - \theta = \underbrace{T_n - E(T_n)}_{\text{Fluctuations de } T_n \text{ autour de sa moyenne}} + \underbrace{E(T_n) - \theta}_{\text{Erreur systématique}}$$



➢ La constante $b = E(T_n) - \theta$ est appelé le <u>biais</u> de l'estimateur et on dit que <u>l'estimateur est sans biais</u> si $E(T_n) = \theta$.

Exemple : Si l'échantillon suit une loi $\mathcal{P}(\lambda)$ alors la moyenne est un estimateur sans biais de λ.

QUALITE D'UN ESTIMATEUR

Convergence de l'estimateur

$\Leftrightarrow$ Améliorer l'estimation quand la taille de l'échantillon augmente

- La loi des grands nombres permet dans certains cas d'établir la convergence en probabilité

$$T_n \xrightarrow[n\to\infty]{p} \theta$$

- La précision de l'estimateur est mesurée par l'erreur quadratique moyenne :

$$R_\theta(T_n) = E[(T_n-\theta)^2] = var(T_n) + b^2$$

Il faut que la précision augmente avec la taille de l'échantillon, *i.e* $R_\theta(T_n) \xrightarrow[n\to\infty]{} 0$

L'estimateur T_1 est meilleur que l'estimateur T_2 si

$$E[(T_1-\theta)^2] \leq E[(T_2-\theta)^2] \quad \Leftrightarrow \quad R_\theta(T_1) \leq R_\theta(T_2)$$

Si estimateurs sans biais, le meilleur est celui de variance minimale

Exemple : Dans le cas où l'échantillon suit une loi de $\mathcal{P}(\lambda)$ alors la moyenne converge en probabilité vers λ et le risque quadratique converge vers 0. On note que la moyenne est un meilleur estimateur de λ que l'estimateur donné par X_1

QUALITE D'UN ESTIMATEUR

Loi asymptotique de l'estimateur

- Le TCL permet dans certains cas d'établir la convergence en loi de l'estimateur

$$T_n \xrightarrow[n\to\infty]{\mathcal{L}} N(*,*)$$

Connaître la loi de l'estimateur permet de calculer des probabilités sur l'estimateur

Exemple : Dans le cas où l'échantillon suit une loi de $\mathcal{P}(\lambda)$ alors l'estimateur donné par la moyenne converge en loi

$$\sqrt{n}\frac{\overline{X}-\lambda}{\sqrt{\lambda}} \xrightarrow[n\to+\infty]{\mathcal{L}} N(0,1)$$

On en déduit par exemple la probabilité d'une précision fixée par ε

$$P\left(\left|\overline{X}-\lambda\right|\leq\varepsilon\right)=2F(\varepsilon\sqrt{\frac{n}{\lambda}})-1$$

où F est la fonction de répartition de la loi N(0,1)

ESTIMATEUR USUEL

La fréquence

Soit un événement A tel que P(A)=p. Sur un échantillon considérons les v.a. $X_1,\ldots,X_n$ telles que $X_i = \begin{cases} 1 & \text{si A est réalisé} \\ 0 & \text{sinon} \end{cases}$. Alors la fréquence (empirique) de l'événement A est définie par

$$\boxed{F_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{K}{n}}$$

où K représente le nombre de fois où l'événement A s'est réalisé au cours des n expériences.

- sans biais : $E(F_n) = p$
- converge en m.q. tel que $R_p(F_n) = \dfrac{p(1-p)}{n}$
- converge en probabilité
- loi asymptotique : $F_n \xrightarrow[n\to+\infty]{\mathcal{L}} N\left(p, \dfrac{p(1-p)}{n}\right)$

Exemple : Une machine fabrique des pièces et le cahier des charges de la machine précise un taux de 10% de pièces défectueuses. On effectue un contrôle. Si la machine est en bon état de fonctionnement, combien faut-il vérifier de pièces pour avoir un taux de pièces défectueuses observé inférieur à 15% avec une probabilité de 95%?

ESTIMATEUR USUEL

La moyenne

La moyenne (empirique) de l'échantillon $X_1,\ldots,X_n$ est la v.a. définie par

$$\boxed{\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i}$$

Rq. La fréquence est une moyenne d'un échantillon de Bernoulli

Supposons que $E(X_i)=\mu$ et $\mathrm{var}(X_i)=\sigma^2$, alors

- sans biais : $E(\overline{X})=\mu$
- converge en m.q. tel que

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

- converge en probabilité
- loi asymptotique :

$$\overline{X}\xrightarrow[n\to+\infty]{\mathcal{L}}N\left(\mu,\frac{\sigma^2}{n}\right)$$

Exemple : On prélève 40 pièces dans une production industrielle. Le cahier des charges de la machine stipule que les pièces sont produites avec un diamètre d'espérance 20 mm et un écart-type de 3 mm. Toutes les pièces ayant un diamètre moyen supérieur à 21 ou inférieur à 19 sont inutilisables. Quel est le pourcentage de pièces prélevées utilisables?

ESTIMATEUR USUEL

La variance empirique

La variance empirique de l'échantillon $X_1,\dots,X_n$ est la v.a. définie par

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

Supposons que $E(X_i)=\mu$ et $var(X_i)=\sigma^2$, alors

- biaisé : $E(S^2) = \frac{n-1}{n}\sigma^2$

$\Rightarrow \boxed{S^{*2} = \frac{n}{n-1}S^2}$ sans biais

- converge en probabilité

- risque quadratique et loi asymptotique inconnus

Exemple : On prélève 40 pièces dans une production industrielle. Le cahier des charges de la machine stipule que le diamètre des pièces suit une loi normale d'espérance 20mm et d'écart-type 2 mm. On considère que la production n'est plus homogène (disparité importante entre les pièces) lorsque l'écart-type empirique dépasse de 5% sa valeur théorique. Quelle est la probabilité que la production ne soit plus homogène ?

ESTIMATEUR USUEL

Echantillon gaussien

Dans le cas d'un échantillon $X_1,\dots,X_n$ gaussien, *i.e* $X_i \sim N(\mu,\sigma^2)$, on connaît la loi exacte des estimateurs usuels de μ et σ^2. Il est donc possible de travailler avec des échantillons de petite taille

Estimation de μ

- σ^2 connu, $\theta=\mu$ $\quad \Rightarrow \quad \sqrt{n}\dfrac{\overline{X}-\theta}{\sigma} \sim N(0,1)$
- σ^2 inconnu, $\theta=\mu$ $\quad \Rightarrow \quad \sqrt{n}\dfrac{\overline{X}-\theta}{S^*} \sim t_{(n-1)}$ (Loi de Student)

Estimation de σ^2

- μ connu, $\theta=\sigma^2$ $\quad \Rightarrow \quad \dfrac{1}{\theta}\sum_{i=1}^{n}(X_i-\mu)^2 \sim \chi^2_n$ (Loi du chi-deux)
- μ inconnu, $\theta=\sigma^2$ $\quad \Rightarrow \quad \dfrac{1}{\theta}\sum_{i=1}^{n}(X_i-\overline{X})^2 \sim \chi^2_{n-1} \Leftrightarrow \dfrac{(n-1)}{\theta}S^{*2} \sim \chi^2_{n-1}$

On sait aussi que la variance empirique S^{*2} converge en m.q.

Exemple : Une machine fabrique des pièces dont le diamètre suit une loi normale $N(20,2^2)$. On prélève 9 pièces dans la production.
Quelle est la probabilité pour que le diamètre moyen des 9 pièces soit inférieur à 18?
Quelle est la probabilité pour que l'écart-type soit supérieur à 2,5?

Estimation par I.D.C.

Estimateur d'un paramètre θ = variable aléatoire

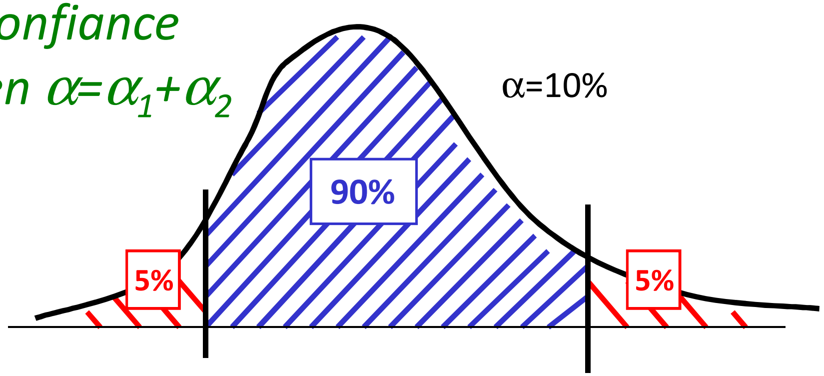
Incertitude sur la valeur estimée

<u>Intervalle de confiance</u> est un intervalle dans lequel on peut supposer que le paramètre θ se trouve avec une probabilité d'erreur fixée par l'utilisateur :

$$P[a \leq \theta \leq b] = 1-\alpha$$

où α est le niveau de confiance

Les bornes de l'intervalle de confiance dépendent du partage de α en $\alpha=\alpha_1+\alpha_2$



- <u>Intervalle bilatéral</u> : [a,b].
  On utilise en général le **risque symétrique**, *i.e*
  $\alpha_1=\alpha_2=\alpha/2$
- <u>Intervalle unilatéral</u>
  - [a,+∞[(*i.e.* $\alpha_1=\alpha$, $\alpha_2=0$) utiliser dans le cas où le paramètre θ doit vérifier $\theta \geq a$ (durée de vie, résistance à la rupture, etc…)
  -]-∞,b[(*i.e.* $\alpha_1=0$, $\alpha_2=\alpha$) utiliser dans le cas où le paramètre θ doit vérifier $\theta \leq b$ (nombre de pièces défectueuses, temps d'attente, etc…)

Estimation par I.D.C.

Méthodologie

- Définir un estimateur $T_n = f(X_1,\ldots,X_n)$ de θ
- Déterminer la loi de l'estimateur
- Transformer l'équation

$$P[a \leq \theta \leq b] = 1-\alpha \quad \Leftrightarrow \quad P[a' \leq T_n \leq b'] = 1-\alpha$$

- Trouver a' et b' grâce à la table de la loi de l'estimateur

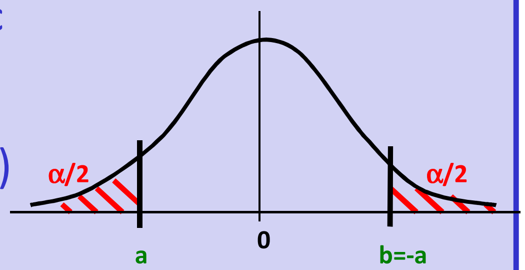
(1 équation et 2 inconnues $\Rightarrow$ faire un dessin pour simplifier le problème)

- En déduire a et b

Exemple : Supposons $X \sim N(\mu,1)$. On cherche un intervalle de confiance pour μ à 95%.

$\overline{X}$ est un estimateur de μ et on sait que $\sqrt{n}(\overline{X}-\mu) \sim N(0,1)$ donc

$$P[a \leq \mu \leq b] = P[\sqrt{n}(\overline{x}-b) \leq \sqrt{n}(\overline{X}-\mu) \leq \sqrt{n}(\overline{x}-a)]$$
$$= F\left(\sqrt{n}(\overline{x}-a)\right) - F\left(\sqrt{n}(\overline{x}-b)\right) = F(a') - F(b')$$



Risque symétrique et loi N(0,1) symétrique /r à Oy $\Rightarrow$ a'=-b'

$$P[a \leq \mu \leq b] = 0.95 \Leftrightarrow 2F(a') - 1 = 0.95 \Leftrightarrow F(a') = 0.975 \Leftrightarrow a' = 1.96$$

$$\Rightarrow a = \overline{x} - \frac{a'}{\sqrt{n}} = \overline{x} - \frac{1.96}{\sqrt{n}} \qquad \text{et} \qquad b = \overline{x} + \frac{a'}{\sqrt{n}} = \overline{x} + \frac{1.96}{\sqrt{n}}$$

Attention de ne pas confondre l'IDC, $P[a \leq \theta \leq b] = 1-\alpha$, et l'intervalle de probabilité $P[a \leq T_n \leq b] = 1-\alpha$

Estimation par I.D.C.

Taille de l'échantillon

Le calcul d'un I.D.C. [a,b] tel que $P[a \leq \theta \leq b] = 1-\alpha$ dépend

- de l'estimateur de θ ($\overline{X}$, S^2,…)
- des quantiles de la loi (obtenus dans les tables)
- de la taille de l'échantillon n

Il est alors possible de déterminer la taille n telle que la longueur de l'intervalle $\ell = b-a$ vérifie une propriété donnée.

Exemple : Un laborantin utilise un thermomètre sophistiqué mais dont les mesures sont incertaines. On sait que l'écart-type de l'appareil est de 1.1 degré Celsius. Combien doit-il effecteur de mesures pour connaître avec un coefficient de confiance de 95%, la température moyenne à 0.5 degré près ?