TP19/20 : Estimation ponctuelle et notion d'intervalle de confiance

▶ Dans votre dossier `Info_2a`, créez le dossier `TP_19`.

I. Quelques rappels de cours

I.1. Contexte de l'estimation

Reprenons l'exemple du cours : on considère une urne contenant des boules vertes et des boules rouges dont on ignore le nombre et la proportion. On effectue un tirage dans cette urne et on note X la v.a.r. égale à 1 si la boule obtenue est verte et à 0 sinon.

Cet exemple permet de comprendre le contexte de l'estimation.

- On considère un phénomène aléatoire et une v.a.r. X qui lui est liée.
- La loi de X est globalement connue *(X suit une loi de Bernoulli dans l'exemple précédent)* mais pas entièrement spécifiée *(le paramètre p est à déterminer).*
- Le problème de l'estimation consiste alors à estimer la valeur du paramètre θ (non spécifié) de cette loi *(p dans notre exemple).*

I.2. Estimateur

Afin d'estimer le paramètre θ, on observe n fois le phénomène *(on tire n fois dans l'urne).*

- On obtient alors un n-uplet de données appelé **échantillon observé**.
- Cette observation est une **réalisation** de n v.a.r. $X_1, \dots, X_n$ mutuellement indépendantes et de même loi que X et définies sur le même espace probablisable *(X_i est la v.a.r. égale à 1 si le $i^{ème}$ tirage donne une boule verte et à 0 sinon).*

Un estimateur T_n du paramètre θ est une v.a.r. $\varphi(X_1, \dots, X_n)$ fonction de l'échantillon $(X_1, \dots, X_n)$.	Toute réalisation $\varphi(x_1, \dots, x_n)$ est appelé estimation de θ.

I.3. Critères de qualité de l'estimateur

Pour tenter d'évaluer la qualité d'un estimateur on dispose des critères suivants.

- **Estimateur sans biais** : $\boxed{\mathbb{E}_\theta(T_n) = \theta}$ *(défini si T_n admet une espérance)*

  On note $b_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta$ le biais de l'estimateur T_n.
- **Estimateur asymptotiquement sans biais** : $\boxed{\lim\limits_{n\to+\infty} \mathbb{E}_\theta(T_n) = \theta}$
- **Estimateur convergent** : $\boxed{\forall \varepsilon > 0,\ \lim\limits_{n\to+\infty} \mathbb{P}_\theta([|T_n - \theta| \geqslant \varepsilon]) = 0}$

  La notion d'estimateur convergent peut être caractérisée à l'aide du **risque quadratique** :

$$\boxed{\begin{array}{rcl} r_\theta(T_n) & = & \mathbb{E}_\theta((T_n - \theta)^2) \\ & = & \mathbb{V}_\theta(T_n) + (b_\theta(T_n))^2 \end{array}}$$ *(défini si, pour tout θ, T_n^2 admet une espérance)*

On a alors : $\boxed{\lim\limits_{n\to+\infty} r_\theta(T_n) = 0 \ \Rightarrow\ T_n \text{ est un estimateur convergent}}$

II. Estimation du paramètre p d'une loi de Bernoulli

On prend (une nouvelle fois) l'exemple du cours : une urne contient des boules vertes et des boules rouges dont on ignore le nombre et la proportion. On effectue un tirage dans cette urne et on note X la v.a.r. égale à 1 si la boule obtenue est verte et à 0 sinon.

Autrement dit, $X \hookrightarrow \mathcal{B}(p)$ où p est un paramètre à déterminer.

II.1. Aspect théorique

- Quel estimateur T_n peut-on proposer ?

- Calculer le biais de cet estimateur.

- Calculer la variance de cet estimateur.

- En déduire que cet estimateur est convergent.

- Quel résultat permet, sans calcul de $\mathbb{V}(T_n)$, de conclure que $T_n = \overline{X_n}$ est convergent ?

II.2. Modélisation en Scilab

Dans la suite, on considère que l'urne contient 300 boules dont 192 vertes.
Pour estimer la proportion p de boules vertes dans cette urne :

× on effectue n tirages successifs et avec remise,

× on calcule la moyenne de ces observations.
(c'est une réalisation de la moyenne empirique)

▶ Quelle est la valeur exacte de p ?

▶ Quel appel permet de simuler une v.a.r. qui suit une loi de Bernoulli de paramètre `p` ?
Comment simule-t-on `n` v.a.r. indépendantes suivant une loi de Bernoulli de paramètre `p` ?

▶ Comment peut-on calculer la moyenne d'un tableau `Obs` contenant `n` éléments.

▶ Écrire une fonction `estimME` qui prend en paramètre un réel `p` et un entier `n` et qui renvoie, dans une variable `T` une simulation de l'estimateur T_n. On utilisera la fonction `grand`.

▶ En théorie, avec cette méthode, quelle est la proportion minimale que l'on peut observer ? Et la maximale ?

▶ Exécuter 8 fois de suite la fonction `estimME` avec les paramètres `p=0.64` et `n=15`.
Noter ci-dessous les résultats.

▶ Exécuter 8 fois de suite la fonction `estimME` avec les paramètres `p=0.64` et `n=50`.
Noter ci-dessous les résultats.

▶ Commenter les résultats obtenus.

II.3. Garantie offerte par cette méthode

Les résultats précédents sont difficilement exploitables tel quels. En effet, ils dépendent fortement de la manière dont les tirages ont été effectués. A-t-on eu plutôt la main chaude (tendance à obtenir plus de succès que prévu) ou plutôt la main froide (tendance à obtenir plus d'échecs) ? Il est compliqué de s'entraîner à avoir la main tiède. Il nous faut donc un procédé mathématique permettant d'estimer la qualité des résultats obtenus.

II.3.a) Loi approchée de T_n par simulation

Dans le chapitre précédent, on a simulé T_n. Essayons de déterminer la loi de T_n par simulation. Pour ce faire :

× on se fixe un n (`n1 = 15` puis `n2 = 50` et enfin `n3 = 150`),

× on simule N réalisations de la v.a.r. T_n (prendre N grand, `N = 10000` par exemple),

× on trace le diagramme en bâtons de ces N valeurs.

▶ À n fixé (prenons $n = 15$ par exemple), quel est le support de T_n ?

▶ Si `n` est fourni, comment obtient-on le support de $T_{\texttt{n}}$ en **Scilab** ?

▶ Écrire un programme **Scilab** qui :

× réalise les affectations initiales `p = 0.64`, `n1 = 15` et `N = 10000`,

× crée un vecteur ligne `ObsT1` de `N` cases et affecte chaque case à une simulation de $T_{\texttt{n}}$, *(on utilisera une structure itérative et la fonction* `estimME` *définie précédemment)*

× affiche le diagramme en bâtons associé aux valeurs de ce vecteur. *(on utilisera la fonction* `histplot` *et on prendra soin de prendre en paramètre un vecteur de classes cohérent)*

- Par quel appel **Scilab** peut-on simuler `N` observations de `n` v.a.r. de Bernoulli indépendantes et de paramètre `p` ?

- En déduire comment l'on peut (avantageusement ?) remplacer la structure itérative précédente à l'aide de fonctions prédéfinies en **Scilab**.

- Ajouter au programme précédent les variables `n2 = 50` et `n3 = 150` et tracer les diagrammes des vecteurs `ObsT2` et `ObsT3` correspondants.
  Les 3 diagrammes doivent être affichés côte à côte. *(on utilisera la fonction* `subplot`*)*

- Commenter les graphiques obtenus :
  1) quel type de courbe obtient-on ?
  2) quel théorème permettait de prévoir ce résultat à l'avance ?

▶ Où lit-on les éléments caractéristiques (espérance, écart-type) d'une loi normale ?
Déterminer les éléments caractéristiques pour les 3 courbes précédentes (lecture graphique et calcul de la valeur attendue).

▶ En quoi la $3^{\text{ème}}$ expérience fournit-elle des résultats meilleurs que les 2 premières ?

II.3.b) Intervalle de confiance asymptotique : aspect théorique

On cherche maintenant à estimer quelle garantie on peut accorder au résultat précédent. A priori, on peut obtenir toutes les valeurs de $[0, 1]$ par nos tirages. Cependant, certaines valeurs (celles très éloignées de p) n'ont qu'une probabilité très faible d'être obtenues alors que d'autres (qui sont proches de p) ont une probabilité forte d'être obtenues.
Ceci invite à poser la notion de garantie sous la forme :

« le résultat obtenu est ε-proche du résultat réel avec une probabilité d'au moins $1 - \alpha$ »

(si on prend $\alpha = 0,05$, on obtient un niveau de confiance de 95%)

▶ Comment exprime-t-on, à l'aide de l'opérateur $|.|$, le fait que deux réels t et p soient écartés au plus de ε ? À quel encadrement de p cela correspond-il ?

▶ En déduire une expression de la garantie qu'il semble raisonnable d'exiger dans notre exemple.

Cette garantie n'est rien d'autre que la notion d'intervalle de confiance.

Intervalles de confiance asymptotique

Soit $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ.

1) Pour $n \in \mathbb{N}^*$, on dit que $[U_n, V_n]$ est un intervalle de confiance asymptotique de θ au niveau de confiance $1-\alpha$ (où le risque α est un élément de $]0,1[$) si :

$$\forall \theta \in \Theta, \quad \boxed{\mathbb{P}_\theta(U_n \leqslant \theta \leqslant V_n) \geqslant 1-\alpha}$$

(classiquement obtenu par l'inégalité de Bienaymé-Chebychev)

2) On dit que $([U_n, V_n])_{n\in\mathbb{N}^*}$ est un intervalle de confiance asymptotique de θ au niveau de confiance $1-\alpha$ (où le risque α est un élément de $]0,1[$) si :

$$\forall \theta \in \Theta, \quad \boxed{\lim_{n\to+\infty} \mathbb{P}_\theta(U_n \leqslant \theta \leqslant V_n) \geqslant 1-\alpha} \quad (1)$$

(classiquement obtenu par le TCL)

Deux notions entrent ici en compétition :

× la précision de l'intervalle (*i.e.* l'écart $V_n - U_n$),

× et le niveau de confiance $(1-\alpha)$ que l'on peut accorder à cet intervalle.

- Plus la précision exigée augmente, plus le niveau de confiance .
  (une précision importante signifie que l'écart $V_n - U_n$ est)
- Plus la précision exigée diminue, plus le niveau de confiance .
  (une précision faible signifie que l'écart $V_n - U_n$ est)

II.3.c) Intervalle de confiance : retour à notre exemple

▶ Le TCL permet d'affirmer que la suite $\left(\overline{X_n}^*\right)$ converge en loi vers une v.a.r. N telle que $N \hookrightarrow \mathcal{N}(0,1)$. Que cela signifie-t-il sur la suite $\left(\mathbb{P}(-x_0 \leqslant \overline{X_n}^* \leqslant x_0)\right)_n$? (avec $x_0 \in \mathbb{R}$)
(on exprimera le résultat à l'aide de Φ)

(1) La définition du BO est : $\forall \theta \in \Theta,\ \exists (\alpha_n) \in [0,1]^{\mathbb{N}^*},\ \forall n \in \mathbb{N}^*,\ \mathbb{P}_\theta(U_n \leqslant \theta \leqslant V_n) \geqslant 1-\alpha_n$ où (α_n) est une suite qui tend vers α. Cette définition moins stricte vise les cas où la suite $(\mathbb{P}_\theta(U_n \leqslant \theta \leqslant V_n))$ ne serait pas convergente. Cette précaution est donc inutile dans le cas d'IC obtenus par le TCL.

- À quel encadrement de p l'inégalité $-x_0 \leqslant \overline{X_n}^* \leqslant x_0$ est-elle équivalente ?

- En déduire l'intervalle de confiance associé à notre problème.

- Quelle valeur de x_0 (expression en fonction de Φ^{-1}) permet d'assurer un niveau de confiance de $1-\alpha$? Quelle est la largeur de l'intervalle (précision) associée ?

- Quel résultat retrouve-t-on alors ?

En **Scilab**, on peut réaliser de nombreux calculs sur Φ grâce à la commande `cdfnor` (*cumulative distribution function normal distribution*).

▶ Que réalise l'appel `P = cdfnor("PQ",x0,0,1)` ?

▶ Que réalise l'appel `x0 = cdfnor("X",0,1,u,1-u)` ?

▶ En déduire l'appel permettant le calcul de $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.

▶ Quelle valeur de x_0 permet d'assurer un niveau de confiance à 85% ?
Quelle valeur de x_0 permet d'assurer un niveau de confiance à 90% ?
Quelle valeur de x_0 permet d'assurer un niveau de confiance à 95% ?
On déterminera les précisions (largeur de l'intervalle) associées dans le cas où $n = 150$.

II.3.d) Intervalle de confiance asymptotique : visualisation

On fixe maintenant `n = 150`, `alpha = 0.05` (niveau de confiance de 95%).
Afin de visualiser la notion d'intervalle de confiance, on agit comme suit :

× on simule `N = 100` réalisations de T_{150},

× pour chaque réalisation de T_{150}, on obtient une réalisation $[u, v]$ de $[U_{150}, V_{150}]$ associée,

- ***(i)*** si p est dans l'intervalle de confiance $[u, v]$, on trace un trait vert vertical d'ordonnée minimale u et maximale v.
- ***(ii)*** si p n'est pas dans l'intervalle de confiance $[u, v]$, on trace un trait rouge vertical d'ordonnée minimale u et maximale v.

× on trace enfin un trait horizontal pour visualiser la valeur de `p`.

- Compléter le programme suivant afin qu'il réalise la stratégie précédente.

```
p = 0.64
n = 150
N = 100
alpha = 0.05

x0 = cdfnor(                          )

clf()
for i = 1:N
    T = estimME(p, n)
    eps = x0/(2*sqrt(n))
    if                  then
        plot([i, i], [T-eps, T+eps], "g")
    else
        plot([i, i], [T-eps, T+eps], "r")
    end
end
plot([1, N], [p, p])
// "Astuce" pour que les ordonnées soient entre 0 et 1
plot([0, 0], [0, 1], "b")
```

- Commenter le résultat obtenu.

III. Estimation du paramètre a d'une loi uniforme sur $[0, a]$

On considère une v.a.r. X telle que $X \hookrightarrow \mathcal{U}([0, a])$ où a est le paramètre à estimer.
On commence par fixer la valeur de a par l'appel suivant : `a = 350`.

III.1. Estimateur sans biais

Dans un premier temps, on considère l'estimateur T_n défini par : $T_n = \dfrac{2}{n} \sum_{i=1}^{n} X_i$ où (X_n) est une suite de v.a.r. indépendantes identiquement distribuées suivant toutes la loi $\mathcal{U}([0, a])$.

- Démontrer que cet estimateur est sans biais.

- Calculer la variance de cet estimateur.

- En déduire que T_n est convergent.

- Dans un nouvel onglet **SciNotes**, mettre en place cette stratégie d'estimation (on s'inspirera du travail précédent).

III.2. Estimateur avec biais

On considère maintenant l'estimateur T_n défini par : $T_n = \max(X_1, \ldots, X_n)$ où (X_n) est une suite de v.a.r. indépendantes identiquement distribuées suivant toutes la loi $\mathcal{U}([0, a])$.

- Calculer la fonction de répartition de T_n et la densité associée.

- Calculer le biais de cet estimateur.

- Montrer que la variance de cet estimateur tend vers 0.

- Peut-on en conclure que T_n est convergent ?

- Dans un nouvel onglet **SciNotes**, mettre en place cette stratégie d'estimation (on s'inspirera du travail précédent).
- Comment pourrait-on corriger la valeur de T_n afin d'obtenir un estimateur sans biais ?

- Démontrer que l'estimateur obtenu est convergent.