TP Informatique n° 14/15
Estimation ponctuelle et intervalle de confiance

I. Aspects théoriques

I.1. Contexte de l'estimation

Dans la suite, on considère l'exemple suivant : 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. Le but est d'estimer la proportion p de boules vertes.

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** : $\mathbb{E}_\theta(T_n) = \theta$ *(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** : $\lim\limits_{n\to+\infty} \mathbb{E}_\theta(T_n) = \theta$
- **Estimateur convergent** : $\forall \varepsilon > 0,\ \lim\limits_{n\to+\infty} \mathbb{P}_\theta(|T_n - \theta| \geqslant \varepsilon) = 0$

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

$$\begin{array}{rcl} r_\theta(T_n) & = & \mathbb{E}_\theta((T_n - \theta)^2) \\ & = & \mathbb{V}_\theta(T_n) + (b_\theta(T_n))^2 \end{array}$$

(défini si, pour tout θ, T_n^2 admet une espérance)

On a alors : $\lim\limits_{n\to+\infty} r_\theta(T_n) = 0 \quad \Rightarrow \quad T_n$ 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 Python

- 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)
- Pour les simulations, on pourra utiliser la fonction `binomial`, accessible depuis l'espace de nommage `numpy.random` et qui permet de simuler une v.a.r. suivant une loi binomiale.
  Plus précisément, `numpy.random(n,p,(nbL,nbC))` renvoie un tableau de taille `nbL` × `nbC` dont chaque case est la simulation d'une v.a.r. X telle que $X \hookrightarrow \mathcal{B}(\mathtt{n}, \mathtt{p})$.
- On commencera par l'importation habituelle.

```
1 import numpy as np
2 import matplotlib.pyplot as plt
```

► 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 une simulation de l'estimateur T_n.

► 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 **Python** ?

- Écrire un programme **Python** qui :
  - réalise les affectations initiales `p = 0.64`, `n1 = 15` et `N = 10000`,
  - crée une liste `ObsT1` de `N` simulations de $T_{\texttt{n}}$,
  - affiche le diagramme en bâtons associé aux valeurs de ce vecteur.
    (on utilisera la fonction `hist` *et on prendra soin de prendre en paramètre un nombre de classes cohérent)*

- 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 et 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 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 :

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

(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)

- 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 ?

La fonction Φ^{-1}, inverse de la fonction de répartition Φ, est appelée fonction **quantile** (parfois appelée *percent point function* en anglais). En **Python**, cette fonction est accessible dans l'espace de nommage `norm` qu'il faut importer depuis la bibliothèque `scipy.stats`.

```
from scipy.stats import norm
```

L'appel $\Phi^{-1}(v)$ est réalisé en **Python** par la commande `norm.ppf(v)`

▶ 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 = norm.ppf(1-alpha/2)

for i in range(N) :
    T = estimME(p, n)
    eps = x0 / (2*np.sqrt(n))
    if                  :
        plt.plot([i+1, i+1], [T-eps, T+eps], 'g')
    else :
        plt.plot([i+1, i+1], [T-eps, T+eps], 'r')
plt.plot([1, N], [p, p])
# "Astuce" pour que les ordonnées soient entre 0 et 1
plt.plot([0, 0], [0, 1], "b")
```

► Commenter le résultat obtenu.

III. Test de conformité à la moyenne

III.1. Introduction

- On considère n réalisations $(x_1, \dots, x_n)$ censées provenir d'un échantillon $(X_1, \dots, X_n)$ de v.a.r. indépendantes, identiquement distribuées et d'espérance m. Le test de conformité à la moyenne consiste à confronter la moyenne de ces réalisations à la moyenne théorique m. L'idée est de pouvoir répondre à la question : les réalisations obtenues sont-elles conformes à nos hypothèses sur les v.a.r. X_i et leur moyenne.
- Dans la pratique, il s'agit souvent de tester si un échantillon est représentatif d'une population.
- Cela peut servir pour des tests de qualité. Considérons par exemple le test de qualité d'une chaîne de production de médicaments. Chaque médicament est censé contenir 12 mg de principe actif. On teste alors un échantillon pour déterminer s'il appartient à la famille des médicaments conformes de la chaîne.
  Plus précisément, on calcule la moyenne de cet échantillon prélevé.
  1) soit le poids moyen du produit actif ne s'écarte pas significativement des 12 mg. Dans ce cas, on rejette l'hypothèse de conformité.
  2) dans le cas contraire, on ne peut rejeter l'hypothèse.

III.2. Rappels théoriques

III.2.a) Si l'on connaît la variance théorique

Dans la section précédente, nous avons montré comment obtenir des intervalles de confiance asymptotique. On a en fait détaillé le théorème suivant (c'est un corollaire immédiat du TCL).

Intervalles de confiance asymptotique via le TCL

- Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a.r. **indépendantes**, identiquement distribuées, d'espérance m et de variance σ.
- Soit $\alpha \in]0,1[$. On note $u_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$. Par définition, $u_{1-\frac{\alpha}{2}}$ est l'unique réel tel que : $\Phi\left(u_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$ *i.e.* $u_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.
- On note $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *(moyenne empirique).*

On a alors : $\lim\limits_{n\to+\infty} \mathbb{P}\left(\overline{X_n} - u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right) = 1-\alpha$

Ou encore : $\lim\limits_{n\to+\infty} \mathbb{P}\left(m \notin \left[\overline{X_n} - u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}},\ \overline{X_n} + u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right]\right) = \alpha$

Ou enfin : $\lim\limits_{n\to+\infty} \mathbb{P}\left(\left|\overline{X_n} - m\right| > u_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right) = \alpha$

On considère dans la suite un n-uplet censé provenir d'un échantillon $(X_1, \ldots, X_n)$ de v.a.r. telles que $X_i \hookrightarrow \mathcal{N}\left(\mu, \sigma^2\right)$ dont on connaît la variance $\sigma^2 = (0.2)^2$.
Le but de l'exercice est de savoir si l'on peut rejeter l'hypothèse H_0 selon laquelle $\mu = 0.64$.

En **Python**, la génération de n réalisations d'un échantillon $(X_1, \ldots, X_n)$ de v.a.r. suivant une même loi normale peut se faire à l'aide de la fonction `rvs`, accessible dans l'espace de nommage `norm`. Plus précsiément, on doit réaliser l'appel :

```
1  norm.rvs(loc, scale, size)
```

où `loc` est la moyenne de la loi normale considérée, `scale` est son écart-type et `size` est la taille de l'échantillon considéré.

- Compléter le programme suivant. Il doit permettre de répondre à la question : « peut-on rejeter l'hypothèse $H_0 : \mu = 0.64$? ».

```
1  # Valeur des paramètres
2  p = 0.64
3  sig = 0.2
4  n = 50
5  alpha = 0.05
6
7  # Quantile d'ordre 1-alpha/2
8  x0 = norm.ppf(1-alpha/2)
9
10 Obs = norm.rvs(         )
11 Xb =
12 # Bornes de l'intervalle de confiance
13 U =
14 V =
15
16 # Test de conformité à la moyenne
17 print(
```

III.2.b) Si l'on ne connaît pas la variance théorique

Dans le cas où la variance théorique σ n'est pas connue, on utilise le théorème suivant (deuxième version du TCL).

Intervalles de confiance asymptotique via le TCL

- Soit $(X_n)_{n\in\mathbb{N}^*}$ une suite de v.a.r. **indépendantes**, identiquement distribuées, d'espérance m et de variance σ.
- Soit $\alpha \in]0,1[$. On note $u_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi $\mathcal{N}(0,1)$. Par définition, $u_{1-\frac{\alpha}{2}}$ est l'unique réel tel que : $\Phi\left(u_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$ *i.e.* $u_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.
- On note $\overline{X_n} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *(moyenne empirique).*
- On note ${S_n}^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X_n})^2 = \dfrac{1}{n}\left(\sum_{i=1}^{n} {X_i}^2\right) - \overline{X_n}^2$ *(variance empirique).*

On a alors : $\lim\limits_{n\to+\infty} \mathbb{P}\left(\overline{X_n} - u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}} \leqslant m \leqslant \overline{X_n} + u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}\right) = 1-\alpha$

Ou encore : $\lim\limits_{n\to+\infty} \mathbb{P}\left(m \notin \left[\overline{X_n} - u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}},\ \overline{X_n} + u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}\right]\right) = \alpha$

Ou enfin : $\lim\limits_{n\to+\infty} \mathbb{P}\left(\left|\overline{X_n} - m\right| > u_{1-\frac{\alpha}{2}}\dfrac{S_n}{\sqrt{n}}\right) = \alpha$

On reprend l'exemple précédent du n-uplet censé provenir d'un échantillon $(X_1,\dots,X_n)$ de v.a.r. telles que $X_i \hookrightarrow \mathcal{N}\left(\mu,\sigma^2\right)$ dont on ne connaît pas ici la variance σ^2. Le but de l'exercice est de savoir si l'on peut rejeter l'hypothèse H_0 selon laquelle $\mu = 0.64$.

▶ Compléter le programme suivant. Il doit permettre de répondre à la question : « peut-on rejeter l'hypothèse $H_0 : \mu = 0.64$? ».

```
# Valeur des paramètres
p = 0.64
sig = np.random.rand()
n = 50
alpha = 0.05

# Quantile d'ordre 1-alpha/2
x0 = norm.ppf(1-alpha/2)

Obs = norm.rvs(         )
Xb =
S2 =
# Test de conformité à la moyenne
print(
```