Chapitre n°13

Estimation

I. Introduction

Ce chapitre utilise bon nombre de résultats préalablement introduits dans les cours de probabilités et de statistiques. Quelle est la différence entre ces deux domaines ?

Elle est fondamentale. Il faudrait sans doute plus que quelques lignes pour la développer mais voici quelques idées. Bien que ces deux domaines s'intéressent à l'étude de phénomènes aléatoires, d'expériences dont on ne peut pas prédire le résultat avec certitude, leur objet d'étude est bien différent.

Lorsqu'il s'agit de faire des probabilités, on fixe un modèle, une expérience, un univers, des évènements, et une mesure de "la chance" qu'ont ces évènements de se produire. On s'intéresse alors aux propriétés des objets étudiés lors de cette expérience, propriétés des lois, espérances et variances de variables aléatoires. La démarche part donc d'un modèle fixé, et les résultats qui en découlent ne laisse que peu de place au doute.

Les statistiques, quant à elle, partent d'observations et cherchent à comprendre le modèle qu'il y a derrière. Dans un premier temps, nous nous sommes intéressés à des statistiques descriptives. Partant d'observation, nous sommes en mesure de définir une moyenne empirique, une variance, une médiane, des quartiles etc., qui vont nous permettent de mieux comprendre la répartition des données observées, de commencer à observer des tendances.
À présent, nous allons introduire les statistiques inférentielles, c'est à dire (d'après la définition du Larousse), *un ensemble des méthodes permettant de formuler en termes probabilistes un jugement sur une population à partir des résultats observés sur un échantillon extrait au hasard de cette population.* Nous allons partir d'observations d'un phénomène aléatoire et faire l'hypothèse que chacune de ces observations est une "réalisation" d'une même loi sous-jacente. Sauf que cette loi, nous ne la connaissons pas, du moins pas en détails...

En effet, que ce soit pour des raisons théoriques ou empiriques, nous sommes parfois amener à supposer que cette loi est de tel ou tel type, et ce sont les paramètres de la loi qui sont inconnus. Prenons quelques exemples :

Exemple :

- À l'approche du second tour d'une élection présidentielle, on interroge une personne au hasard et on note $X = 1$ si elle se prononce pour le candidat A et $X = 0$ si c'est pour le candidat B. X suit une loi de Bernoulli de paramètre $\theta_0 \in [0,1]$ inconnu qui correspond à la proportion de français qui votent pour A.
- On souhaite modéliser le nombre N de voitures se présentant à un péage en une heure. Il s'agit du nombre de réalisations d'un évènement rare sur un grand nombre d'observations. On suppose donc que ce nombre N suit une loi de Poisson. On cherche le paramètre de cette loi.
- On souhaite modéliser la durée de vie T d'un appareil électrique. Ce phénomène étant sans mémoire, il peut se modéliser à l'aide d'une loi exponentielle. Il s'agit de trouver son paramètre.

Notons X la variable égale au résultat de notre expérience aléatoire. On suppose donc que l'on sait de quel type est la loi de X (elle appartient à une famille de lois $(\mathrm{P}_\theta)_{\theta\in\Theta}$ connue), mais elle dépend d'un paramètre θ inconnu appartenant à Θ, l'espace des paramètres. Le paramètre θ peut être réel (Θ est une partie $\mathbb{R}$) ou vectoriel (Θ est une partie de $\mathbb{R}^k$, $k \geq 2$).

Exemple :

Dans les trois cas précédents :

- $X \hookrightarrow \mathcal{B}(\theta), \Theta =]0,1[$;
- $N \hookrightarrow \mathcal{P}(\theta), \Theta =]0,+\infty[$;
- $T \hookrightarrow \mathcal{E}(\theta), \Theta =]0,+\infty[$.

On peut aussi imaginer le cas de $X \hookrightarrow \mathcal{N}\left(m, \sigma^2\right)$ où $\theta = \left(m, \sigma^2\right)$ est inconnu (paramètre vectoriel, $\Theta = \mathbb{R}\times]0, +\infty[$).

L'objectif de la statistique inférentielle est d'estimer la "vraie" valeur θ_0 du paramètre θ à partir de réalisations de la variable X.

Dans toute la suite n est un entier naturel strictement plus grand que 1 et toutes les variables aléatoires sont définies sur un même espaces probabilisable $(\Omega, \mathcal{A})$ et munit d'une famille de loi de probabilité $(\mathrm{P}_\theta)_{\theta\in\Theta}$.

Définition 1.1

Soit X une variable aléatoire de loi P_θ.

- Un n **échantillon** de la loi X est un n-uplet $(X_1, X_2, \ldots, X_n)$ de variables mutuellement indépendantes [a] et suivant la même loi que X.
- Une **observation** (ou échantillon observé) est un n-uplet $(x_1, x_2, \ldots, x_n)$ de réels tel que x_1, x_2,..., x_n sont les valeurs respectivement prises par X_1, X_2, ... X_n. Autrement dit

$$\exists\omega \in \Omega,\ \forall i \in [\![1,\ n]\!],\ X_i(\omega) = x_i.$$

a. Plus précisément indépendantes pour toutes les lois P_θ

Notation [1] Dans le cas où ils existent on note $E_\theta(X)$ et $V_\theta(X)$ l'espérance et la variance de X pour la loi P_θ

Remarque :

Il faut bien distinguer le n-échantillon $(X_1, \ldots, X_n)$, qui est une variable aléatoire (et donc une fonction), de la réalisation $(X_1(\omega), \ldots, X_n(\omega)) = (x_1, .., x_n)$ qui est un élément de $\mathbb{R}^n$. Si par exemple X suit une loi $\mathcal{E}(\theta)$, alors `rd.exponential(1/theta,n)` fournit une réalisation de cet échantillon.

Exemple :

Dans la cas de l'élection présidentielle, on questionne $n = 5$ individus sur leurs intentions de vote et on obtient les résultats suivants (en notant 1 ou 0 selon que le choix se porte sur le candidat A ou B) :

$$1, \quad 0, \quad 0, \quad 1, \quad 0.$$

Ces résultats observés correspondent, pour tout $\theta \in [0, 1]$, à la réalisation d'un 5-échantillon $(X_1, \ldots, X_5)$ P_θ-indépendant de loi mère $\mathcal{B}(\theta)$. Pour tout $1 \leq i \leq 5$, on a $E_\theta\left(X_i\right) = \theta$ et $V_\theta\left(X_i\right) = \theta(1-\theta)$.

II. Estimation ponctuelle

II. 1 Définitions, exemples

Définition 2.1

Soit $(X_1, X_2, \ldots, X_n\ )$ un n-échantillon d'une v.a. X dont la loi dépend d'un paramètre θ, θ appartenant à une partie $\Theta \subset \mathbb{R}$.

1. On appelle **estimateur de** $g(\theta)$ toute variable aléatoire T_n de la forme

$$T_n = \varphi\left(X_1, X_2, \ldots, X_n\right)$$

où φ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}$, éventuellement dépendante de n mais **indépendante de** θ.

2. Soit $\varphi\left(X_1, \ldots, X_n\right)$ un estimateur de $g(\theta)$. Soit $(x_1, \ldots, x_n) \in X_1(\Omega)\times\ldots\times X_n(\Omega)$. On dit que $\varphi\left(x_1, \ldots, x_n\right)$ est une **estimation ponctuelle de** $g(\theta)$**.**
Autrement dit, une estimation ponctuelle de $g(\theta)$ est une réalisation d'un estimateur de $g(\theta)$.

1. En pratique l'indice θ est omis.

Attention:

On insiste ici sur le fait que, bien qu'un estimateur soit dépendant de θ au sens où il est fonction des X_i dont la loi dépend de θ, **son expression ne doit pas faire intervenir** θ.

Définition 2.2

Soit $(X_1, X_2, \dots, X_n)$ un échantillon de la loi X.

- On définit la moyenne empirique par

$$\overline{X_n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{k=1}^{n} X_i$$

- On définit la variance empirique par

$$V_n = \frac{1}{n}\sum_{k=1}^{n}(X_i - \overline{X_n})^2$$

Exemple :

1. Dans le cas de l'élection présidentielle, la moyenne empirique $\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$ est l'estimateur « le plus naturel» de θ. On obtient avec l'échantillon observé que :

$$\overline{X_5}(\omega) = \frac{1}{5}(1+0+0+1+0) = \frac{2}{5} \text{ est une estimation de } \theta$$

On peut envisager bien d'autres estimateurs pour θ, comme par exemple :

$$A_n = \frac{2}{n(n+1)}\sum_{k=1}^{n} kX_k \quad \text{ou} \quad B_n = 0$$

chacun fournissant une autre estimation de θ à partir de notre échantillon observé :

$$A_n(\omega) = \frac{2}{5\times 6}(1+0+0+4+0) = \frac{1}{3} \quad \text{ou} \quad B_n(\omega) = 0$$

2. Si $(X_1, \dots, X_n)$ est un n-échantillon d'une loi $\mathcal{E}(\theta)$, alors $\overline{X_n}$ est un estimateur de $g(\theta) = \frac{1}{\theta}$, et `np.mean(rd.exponential(1/theta,n))` en est une estimation.
3. Si $(X_1, \dots, X_n)$ est un n-échantillon de loi $\mathcal{U}([a,b])$ de paramètre vectoriel inconnu $\theta = (a,b) \in \mathbb{R}^2$, alors :

$$T_n = \max(X_1, \dots, X_n) - \min(X_1, \dots, X_n)$$

est un estimateur de $g(\theta) = b - a$.

Remarques :

R1 – Certains estimateurs visent à estimer non pas le paramètre mais une fonction $g(\theta)$ du paramètre.

R2 – Dans de nombreux cas, on prendra $g = \mathrm{Id}_{\mathbb{R}}$ et on estimera θ directement.

R3 – La définition précédente est très générale, et la notion même d'estimateur reste un peu floue, car des estimateurs on peut en construire une infinité.... La question qui se pose à nous est alors : *Qu'est-ce qu'un bon estimateur ?*

II. 2 Qualité d'un estimateur

Dans cette partie, on introduit quelques notions qui sont à présent hors programme, mais qui sont importantes dans la compréhension de ce chapitre et qui peuvent faire l'objet de questions au concours (sans que les notations et vocabulaires ne soient introduits).

Biais

Définition 2.3

Soit T_n un estimateur de $g(\theta)$. Si T_n admet une espérance, **le biais de l'estimateur** T_n, noté $b_\theta(T_n)$, est le réel défini par :

$$b_\theta(T_n) = \mathbb{E}_\theta(T_n - g(\theta))$$

Vocabulaire :

- Si $b_\theta(T_n) = 0$, on dit que l'estimateur est sans biais, ce qui signifie que $= E_\theta(T_n) = g(\theta)$. Autrement dit si en moyenne l'erreur commise $T_n - g(\theta)$ est nulle.
- Si $\lim\limits_{n\to+\infty} b_\theta(T_n) = 0$, on dit que l'estimateur est asymptotiquement sans biais.

Proposition 2.4

Si T_n admet une espérance, alors : $b_\theta(T_n) = \mathbb{E}_\theta(T_n) - g(\theta)$

Proposition 2.5

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une variable X admettant une espérance m. Alors la moyenne empirique :

$$\overline{X_n} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

est un estimateur sans biais de m.

Exercice 1

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon d'une variable X de loi $\mathcal{U}([0;\theta]); \theta \in \mathbb{R}_+$.

1. Montrer que $\bar{X}_n$ est un estimateur biaisé et θ et préciser $b_\theta\left(\bar{X}_n\right)$.
2. Proposer alors un estimateur V_n de θ sans biais, obtenu comme transformation simple de $\bar{X}_n$.
3. On considère maintenant l'estimateur $M_n = \max(X_1, X_2, \ldots, X_n)$.
   (a) Déterminer la fonction de répartition F de M_n et en déduire une densité f de M_n.
   (b) Montrer alors que
   $$E_\theta(M_n) = \frac{n\theta}{n+1}.$$
   (c) En déduire un estimateur sans biais Z_n à partir de M_n.

Simulation informatique

On cherche à estimer la paramètre θ inconnu d'une loi uniforme $\mathcal{U}([0,\theta])$. Nous avons vu deux estimateurs sans biais de θ :

$$V_n = \frac{2}{n}(X_1 + X_2 + \ldots + X_n), \quad Z_n = \frac{n+1}{n}\max(X_1, X_2, \ldots, X_n)$$

Comparons numériquement : V_{20}, V_{100} et Z_{100}. On tape alors

```
1  theta=rd.exponential(10)#Choix d'un paramètre theta au hasard
2
3  print('Estimation avec V20')
4  for i in range(10):
5      aux=2*np.mean(rd.uniform(0,theta,20))
6      print("V20",aux,'Ecart entre V20 et theta',np.abs(aux-theta))
7
8
9  print('Estimation avec V100')
10 for i in range(10):
11     aux=2*np.mean(rd.uniform(0,theta,100))
12     print("V100",aux,'Ecart entre V100 et theta',np.abs(aux-theta))
13
14 print('Estimation avec Z100')
15 for i in range(10):
16     aux=max(1.01*rd.uniform(0,theta,100))
17     print("Z100",aux,'Ecart entre Z100 et theta',np.abs(aux-theta))
```

Ce qui renvoie

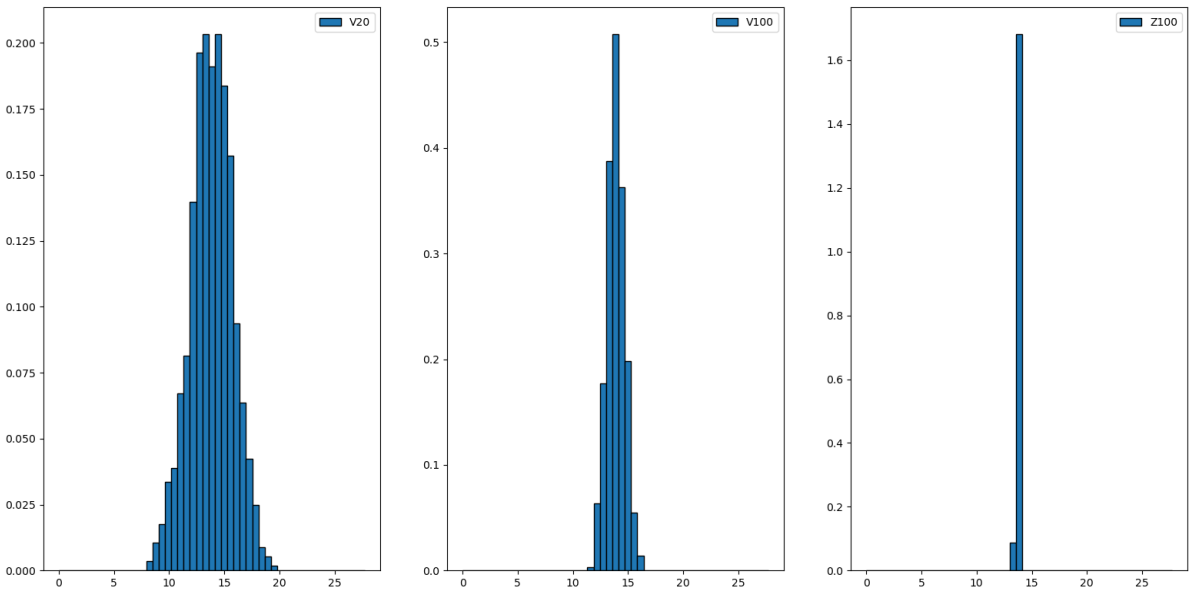
```
>>> (executing cell "Comparaison numér..." (line 8 of "Estimation.py"))
Estimation avec V20
V20 3.4613674019232774 Ecart entre V20 et theta 0.536101038995048
V20 3.138267293789682 Ecart entre V20 et theta 0.21300093086145244
V20 3.238372085102283 Ecart entre V20 et theta 0.31310572217405364
V20 2.8125729936782147 Ecart entre V20 et theta 0.11269336925001472
V20 2.70930245108374 Ecart entre V20 et theta 0.2159639118444896
V20 3.1211729502526415 Ecart entre V20 et theta 0.19590658732441213
V20 2.3837600749956547 Ecart entre V20 et theta 0.5415062879325747
V20 3.4838268727676622 Ecart entre V20 et theta 0.5585605098394328
V20 2.711356019464815 Ecart entre V20 et theta 0.21391034346341442
V20 2.2916323531010256 Ecart entre V20 et theta 0.6336340098272037
Estimation avec V100
V100 2.8682959549247986 Ecart entre V100 et theta 0.056970408003430784
V100 2.9746601277157527 Ecart entre V100 et theta 0.04939376478752333
V100 2.7408614328883205 Ecart entre V100 et theta 0.18440493003990888
V100 3.026503156044636 Ecart entre V100 et theta 0.10123679311640643
V100 2.9474099786381998 Ecart entre V100 et theta 0.022143615709970366
V100 2.606450112093952 Ecart entre V100 et theta 0.31881625083427734
V100 3.1239056205439244 Ecart entre V100 et theta 0.198639257615695
V100 3.0020116784470376 Ecart entre V100 et theta 0.07674531551880825
V100 2.875695828840214 Ecart entre V100 et theta 0.04957053408801526
V100 3.0238892610861603 Ecart entre V100 et theta 0.0986228981579309
Estimation avec Z100
Z100 2.95048540175123 Ecart entre Z100 et theta 0.025219038823000695
Z100 2.8700636285976384 Ecart entre Z100 et theta 0.055202734330590975
Z100 2.8689933650651254 Ecart entre Z100 et theta 0.056272997863104024
Z100 2.923126759140127 Ecart entre Z100 et theta 0.0021396037881022956
Z100 2.855677498165216 Ecart entre Z100 et theta 0.06958886476301318
Z100 2.917009076548008 Ecart entre Z100 et theta 0.008257286380221274
Z100 2.951075021845451 Ecart entre Z100 et theta 0.025808658917221727
Z100 2.9237026676384676 Ecart entre Z100 et theta 0.0015636952897617462
Z100 2.8846335399758427 Ecart entre Z100 et theta 0.04063282295238668
Z100 2.92631567882273 Ecart entre Z100 et theta 0.0010493158945004133
```

On peut constater que ces trois estimateurs sont de bons estimateurs de θ, mais en regardant plus en détails, il semble que les valeurs renvoyées par Z_{100} soient plus proches de θ que celle de V_{100}, elles mêmes plus proches que celles de V_{20}.

Examinons maintenant des résultats graphiques en traçant les histogrammes associés à partir de 1000 simulations de chacune de ces variables.

```
theta=rd.exponential(10)

V20=[]
V100=[]
Z100=[]

for i in range(1000):
    V20.append(2*np.mean(rd.uniform(0,theta,20)))
    V100.append(2*np.mean(rd.uniform(0,theta,100)))
    Z100.append(1.01*max(rd.uniform(0,theta,100)))

x=np.linspace(0,2*theta,100)


plt.subplot(131)
plt.hist(V20,x,edgecolor='k',density=True,label='V20')
plt.legend()

plt.subplot(132)
plt.hist(V100,x,edgecolor='k',density=True,label='V100')
plt.legend()

plt.subplot(133)
plt.hist(Z100,x,edgecolor='k',density=True,label='V100')
plt.legend()

plt.show()
```

Voici l'affichage Python pour $\theta = 13.85779831676147$



On observe que la distribution de M_{100} est plus resserrée autour de la valeur de θ que celle de $2\overline{X_{100}}$, elle-même plus resserrée que celle de $2\overline{X_{20}}$.

Exercice 2

Soit $(X_1, \ldots, X_n)$ un n-échantillon d'une v.a. X admettant pour espérance m et pour variance σ^2.

1. On suppose que m est connu. Montrer que V_n est un estimateur non biaisé de σ^2, où $T_n = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$.
2. On suppose que m n'est pas connu. On note $\bar{X}_n$ la moyenne empirique de l'échantillon et V_n la variance empirique
   (a) Montrer que, pour tout $i \in [\![1; n]\!]$, $E\left(\left(X_i - \bar{X}_n\right)^2\right) = V\left(X_i - \bar{X}_n\right)$.
   (b) Montrer que, pour tout $i \in [\![1; n]\!]$, $V\left(X_i - \bar{X}_n\right) = \left(1 - \dfrac{1}{n}\right)^2 V(X_i) + \dfrac{1}{n^2}\sum_{k \neq i} V(X_k)$.
   (c) En déduire que $V\left(X_i - \bar{X}_n\right) = \dfrac{n-1}{n}\sigma^2$..
   (d) Montrer alors que U_n est un estimateur biaisé de σ^2. En déduire un estimateur sans biais de σ^2.

Risque quadratique

Définition 2.6

Soit T_n un estimateur de $g(\theta)$. Si T_n admet une variance, **le risque quadratique de l'estimateur** T_n, noté $r_\theta(T_n)$, est le réel défini par :

$$r_\theta(T_n) = \mathbb{E}_\theta\left((T_n - g(\theta))^2\right)$$

Proposition 2.7

Si T_n admet une variance, alors : $r_\theta(T_n) = \mathbb{V}_\theta(T_n) + (b_\theta(T_n))^2$.

Démonstration. □

Remarque :

> Si T_n est sans biais, alors $r_\theta(T_n) = \mathbb{V}_\theta(T_n)$.

Vocabulaire :
Si T_n et U_n sont deux estimateurs de $g(\theta)$, on dit que T_n *est meilleur que* U_n si et seulement si $r_\theta(T_n) \leq r_\theta(U_n)$ pour tout $\theta \in \Theta$.

Dans la mesure du possible, on cherche à construire un estimateur dont le biais et le risque quadratique sont les plus proches de 0 possible, l'idéal étant un estimateur sans biais et de variance minimale... Mais celui-ci n'existe pas toujours, ou n'est pas simple à trouver. De façon générale, pour comparer deux estimateurs :

- s'ils ont même biais, on préférera celui de risque quadratique minimal. D'après la proposition précédente cela signifie qu'on prendra celui qui a la variance la plus petite.
  - si sa variance est grande, il prendra souvent des valeurs très éloignées de $g(\theta)$;
  - à l'inverse, si sa variance est proche de 0 , il donnera des estimations en moyenne très regroupées autour de $g(\theta)$.
- sinon, on pourra parfois préférer un estimateur biaisé à un estimateur sans biais si son risque quadratique est plus faible que la variance de l'estimateur sans biais.

Exercice 3

On reprend les notations de l'exercice 1. C'est à dire que X suit une loi $\mathcal{U}([0;\theta]); \theta \in \mathbb{R}_+$ et on a les deux estimateurs sans biais de θ

$$V_n = \frac{2}{n}\left(X_1 + X_2 + \ldots + X_n\right), \quad Z_n = \frac{n+1}{n}\max\left(X_1, X_2, \ldots, X_n\right)$$

1. Établir que $r_\theta\left(V_n\right) = V_\theta\left(V_n\right) = \dfrac{\theta^2}{3n}$.
2. Montrer que $E_\theta\left(Z_n^2\right) = \dfrac{(n+1)^2}{n^2}\displaystyle\int_0^\theta \frac{nt^{n+1}}{\theta^n}\,\mathrm{d}t$. En déduire que $r_\theta\left(Z_n\right) = V_\theta\left(Z_n\right) = \dfrac{\theta^2}{n(n+2)}$.
3. Quel estimateur aura-t-on tendance à préférer en pratique ?

Interprétation graphique :
En observant les histogrammes obtenus précédemment, on remarque qu'ils sont "centrés en θ, ce qui signifie que les observations obtenues de $V100$ et $Z100$ sont, en moyenne, égales à θ. Ceci n'a rien d'étonnant puisque les deux estimateurs sont sans biais. Le choix entre ces deux estimateurs va donc reposer sur la dispersion des valeurs par rapport à la moyenne et on retrouve que les valeurs de $V100$ sont plus dispersées que celle de $Z100$, ce qui fait de $Z100$ un meilleur estimateur.

Estimateur convergent

Définition 2.8

Un estimateur T_n de $g(\theta)$ est dit convergent si, pour tout $\theta \in \Theta$,

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} P_\theta\left(|T_n - g(\theta)| > \varepsilon\right) = 0$$

Autrement dit, si on augmente la taille n de l'échantillon, la probabilité que l'estimateur s'écarte de la quantité qu'il estime diminue.

Proposition 2.9

Soit T_n un estimateur tel quel, pour tout $\theta \in \Theta$,

$$\left(\lim_{n\to+\infty} r_\theta\left(T_n\right) = 0\right) \Rightarrow \left(T_n \text{est un estimateur convergent. }\right)$$

Démonstration. Soient $\theta \in \Theta$ et $n \in \mathbb{N}^*$. D'après l'inégalité de Markov, pour tout $\varepsilon > 0$, on a

$$P\left(|T_n - g(\theta)| > \varepsilon\right) = P\left(\left|(T_n - g(\theta))^2\right| > \varepsilon^2\right) \leq \frac{r_\theta\left(T_n\right)}{\varepsilon^2} \underset{n\to+\infty}{\longrightarrow} 0$$

□

Un estimateur sans biais garantit l'absence d'erreur en moyenne, mais peut produire de très mauvaises estimations ponctuelles. Au contraire, un estimateur convergent, qu'il soit biaisé ou non, sera d'autant plus fiable que la taille de l'échantillon est grande. En général, on préfère ces derniers.

Retenir

La loi faible des grands nombres stipule que pour (X_n) une suite de v.a. (mutuellement) indépendantes, admettant une même espérance m et une même variance σ^2, alors

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} P\left(\left|\overline{X}_n - m\right| \geq \varepsilon\right) = 0$$

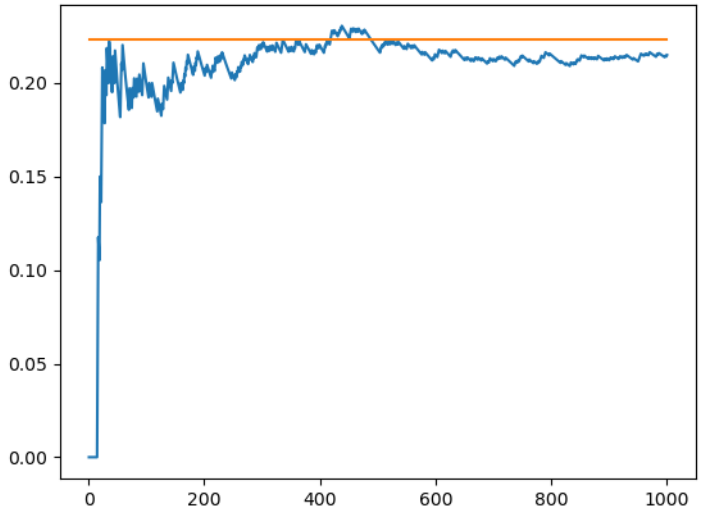
Ceci implique notamment que $\bar{X}_n$ d'un n-échantillon est un estimateur sans biais et convergent de l'espérance.

Nous pouvons illustrer ceci sur Python. Considérons X une variable aléatoire suivant une loi $\mathcal{B}(p)$. Soit $(X_1, X_2, ..., X_n)$ un n-échantillon de X. On sait que $\overline{X_n} = \dfrac{1}{n}\sum_{k=1}^{n} X_i$ est un estimateur de p. Pour mettre en évidence la convergence de $(X_n)_{n\geq 1}$ vers p, nous allons créer le vecteur

$$\left(X_1, \frac{X_1+X_2}{2}, \frac{X_1+X_2+X_3}{3}, ..., \frac{X_1+...+X_n}{n}\right)$$

```
45 p=rd.rand()
46 n=int(input('Entrer un entier n'))
47 x=np.arange(1,n+1)
48 y=rd.binomial(1,p,n)#Simulation de n var de Bernoulli de paramètre p
49 m=np.cumsum(y)/x
50 plt.plot(x,m,x,p*np.ones(n))
51 plt.show()
```

On obtient la représentation suivante, où $p = 0.22338491070969135$ et $n = 1000$:



II. 3 Maximum de Vraisemblance

D'après ce que l'on vient de voir, la loi faible des grands nombres fournit un estimateur de qualité raisonnable de l'espérance d'une loi. Cependant les paramètres à estimer ne sont pas tous des espérances, et lorsque c'est le cas la moyenne empirique n'est pas toujours le meilleur estimateur. Il faut nous doter d'une méthode permettant de déterminer des estimateurs pertinents.

On suppose que l'on dispose d'un n-échantillon $(X_1, X_2, ..., X_n)$ d'une variable X suivant une loi discrète $\mathcal{L}(\theta)$, et d'une observation $(x_1, x_2, ..., x_n)$. Autrement dit on sait que l'évènement

$$[X_1 = x_1] \cap [X_2 = x_2] \cap ... \cap [X_n = x_n]$$

s'est réalisé.
On cherche ici à estimer θ. Il est raisonnable de penser que la valeur du paramètre est celle qui donne la valeur maximale à la probabilité de l'évènement précédent. Autrement dit une estimation ponctuelle de θ viendrait de la valeur maximale de

$$\theta \mapsto \mathrm{P}_\theta\left([X_1 = x_1] \cap [X_2 = x_2] \cap ... \cap [X_n = x_n]\right)$$

ou encore, par indépendance,

$$\theta \mapsto \prod_{i=1}^{n} \mathrm{P}_\theta\left(X_i = x_i\right))$$

Ceci motive les définitions suivantes

Définition 2.10

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu et $(X_1 \dots, X_n)$ un n-échantillon de X. Soit $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$ un n-uplet d'observations.

- Si X est discrète, on appelle **fonction de vraisemblance pour** $(x_1, \dots, x_n)$ la fonction :

$$L_n : \theta \longmapsto \prod_{k=1}^{n} \mathbb{P}_\theta \left([X_k = x_k]\right)$$

- Si X est à densité, de densité f_θ, on appelle **fonction de vraisemblance pour** $(x_1, \dots, x_n)$ la fonction :

$$L_n : \theta \longmapsto \prod_{k=1}^{n} f_\theta \left(x_k\right)$$

L'idée est alors d'attribuer à θ la valeur, si elle existe et si elle est unique, du réel θ_n^* qui maximise la fonction L_n.

Définition 2.11

Soit X une v.a. suivant une loi dépendant d'un certain paramètre θ inconnu et $(X_1 \dots, X_n)$ un n-échantillon de X. Soit $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$ un n-uplet d'observations. On note L_n la fonction de vraisemblance associée et on suppose qu'elle admet un maximum sur Θ, atteint en un unique réel.

- **L'estimation du maximum de vraisemblance de** θ est le réel θ_n^* défini par : $\theta_n^* = \max\limits_{\theta \in \Theta} \left(L_n(\theta)\right)$.
- Si g est la fonction telle que $\theta_n^* = g\left(x_1, \dots, x_n\right)$, alors l'estimateur du maximum de vraisemblance de θ est la variable aléatoire $T_n = g\left(X_1, \dots, X_n\right)$.

Remarque :

Il est plus commode de dériver une somme qu'un produit. On travaille souvent sur la log-vraisemblance, égale à $\ln(L_n)$. Les fonctions vraisemblance et log-vraisemblance ont mêmes variations et mêmes positions d'extrema.

Exemple :

On souhaite déterminer l'estimateur du maximum de vraisemblance du paramètre p d'une loi de Bernoulli. Pour cela, on considère une variable aléatoire X suivant la loi $\mathcal{B}(p)$, avec $p \in]0;1[$ inconnu. Considérons également, pour tout $n \in \mathbb{N}^*$, $(X_1, \dots, X_n)$ un n-échantillon de X et $(x_1, \dots, x_n)$ une réalisation de cet échantillon.

Soit $n \in \mathbb{N}^*$. On a :

$$\forall k \in [\![1;n]\!], \forall x_k \in \{0;1\}, \mathbb{P}\left([X_k = x_k]\right) = \left\{ \begin{array}{ll} p & \text{si } x_k = 1 \\ 1-p & \text{si } x_k = 0 \end{array} \right. = p^{x_k}(1-p)^{1-x_k}$$

Ainsi, en notant L_n la fonction de vraisemblance pour $(x_1, \dots, x_n)$ et $\varphi_n(p) = \ln\left(L_n(p)\right)$, on a, pour tout $p \in]\,0\,;\,1[$:

$$L_n(p) = \prod_{\substack{k=1 \\ n}}^{n} \mathbb{P}\left([X_k = x_k]\right) = \prod_{k=1}^{n} p^{x_k}(1-p)^{1-x_k} \qquad \text{donc } \forall p \in]0;1\,[, \quad L_n(p) > 0$$

et, pour tout $p \in]0;1[$:

$$\begin{aligned} \varphi_n(p) = \ln\left(L_n(p)\right) &= \ln\left(\prod_{k=1}^{n} p^{x_k}(1-p)^{1-x_k}\right) && \forall k \in [\![1;n]\!], p^{x_k}(1-p)^{1-x_k} > 0 \\ &= \sum_{k=1}^{n} \ln\left(p^{x_k}(1-p)^{1-x_k}\right) = \sum_{k=1}^{n} \left(x_k \ln(p) + (1-x_k)\ln(1-p)\right) && p > 0, 1-p > 0 \\ &= \left(\sum_{k=1}^{n} x_k\right)\ln(p) + \left(n - \sum_{k=1}^{n} x_k\right)\ln(1-p) = s_n \ln(p) + (n-s_n)\ln(1-p) && \text{en notant } s_n = \sum_{k=1}^{n} x_k \end{aligned}$$

La fonction φ_n est donc dérivable sur $]0;1[$ et, pour tout $p \in]0;1[$.

$$\varphi_n'(p) = s_n \frac{1}{p} - (n - s_n)\frac{1}{1-p} = \frac{s_n(1-p) - (n-s_n)\,p}{p(1-p)} = \frac{s_n - np}{p(1-p)}$$

Donc la dérivée s'annule en $p = \dfrac{s_n}{n}$, est positive avant, négative ensuite.

La fonction φ_n admet donc un maximum en $\dfrac{s_n}{n} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$.

Or, par stricte croissance de ln sur $\mathbb{R}_*^+$, la fonction φ_n et L_n ont les mêmes variations sur $]0\,;\,1[$.

Par conséquent : la fonction L_n admet un maximum en $\dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$.

Conclusion : l'estimateur du maximum de vraisemblance de p est $\dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k = \overline{X_n}$.

Exercice 4

On considère un n-échantillon $(X_1, .., X_n)$ d'une loi de Poisson de paramètre $\lambda > 0$ inconnu que l'on cherche à estimer, ainsi que $(x_1, .., x_n)$ un n-uplet de $\mathbb{N}^n$ fixé.

1. En notant $s_n = \displaystyle\sum_{i=1}^{n} x_i, \quad p_n = \prod_{i=1}^{n} (x_i!)$, exprimer la fonction de vraisemblance L_n de la loi de Poisson, définie sur $\mathbb{R}_+^*$.
2. Calculer $L_n'(\theta)$ pour tout $\theta > 0$ et vérifier que L_n est maximale en
$$\theta^* = \frac{s_n}{n}$$
3. Quelle est donc l'expression de l'estimateur de maximum de vraisemblance pour la loi de Poisson ?

Exercice 5

On considère un n-échantillon de la loi $\mathcal{U}([0;\theta])$ et on cherche à estimer $\theta > 0$.
Soit $(x_1, ..;, x_n) \in \left(\mathbb{R}_+^*\right)^n$ fixé. On note f_θ la densité de notre loi uniforme. On introduit la fonction de vraisemblance, définie sur $\mathbb{R}_+^*$ par $L_n(\theta) = \displaystyle\prod_{i=1}^{n} f_\theta(x_i)$.

1. Montrer que, pour tout $\theta \geq 0$, on a $L_n(\theta) = \begin{cases} \theta^{-n}, \text{ si } \theta \geq \max(x_1, \ldots, x_n) \\ 0, \text{ sinon} \end{cases}$
2. En déduire que l'estimateur du maximum de vraisemblance pour la loi $\mathcal{U}([0;\theta]$ est donné par
$$\hat{\theta}_n = \max(X_1, \ldots, X_n)$$

III. Estimation par intervalle de confiance

III. 1 Intervalle de confiance (exact)

Définition 3.1

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu. Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ telles que : $\forall n \in \mathbb{N}^*, \mathbb{P}_\theta([U_n \leq V_n]) = 1$.
Soit $\alpha \in]0;1[$. On dit que $[U_n, V_n]$ est **un intervalle de confiance de θ au niveau de confiance $1-\alpha$ (ou au risque α)**, lorsque :

$$\mathbb{P}_\theta([\theta \in [U_n, V_n]]) \geq 1 - \alpha$$

Remarques :

R1 – Les bornes de l'intervalle de confiance ne doivent jamais dépendre du paramètre à estimer.

R2 – Il faut bien identifier que θ n'est pas une variable aléatoire, tandis que l'intervalle $[U_n, V_n]$ est aléatoire.

Exercice 6

On considère une variable aléatoire X suivant une loi $\mathcal{N}\left(\mu;\sigma^2\right)$, où m est inconnu et σ^2 connu non nul. On dispose d'un n-échantillon $(X_1,\ldots,X_n)$ de X

1. Donner la loi de $\overline{X_n}$ et celle de $\bar{X}_n{}^* = \sqrt{n}\dfrac{\overline{X_n}-m}{\sigma}$.
2. Soit $\alpha \in]0;1[$. Justifier l'existence d'un unique réel t_α tel que $\Phi\left(t_\alpha\right) = 1-\dfrac{\alpha}{2}$, où Φ désigne la fonction de répartition d'une variable aléatoire suivant la loi $\mathcal{N}(0;1)$.
3. En déduire que $\left[\overline{X_n}-t_\alpha\dfrac{\sigma}{\sqrt{n}}, \overline{X_n}+t_\alpha\dfrac{\sigma}{\sqrt{n}}\right]$ est un intervalle de confiance de m au niveau de confiance $1-\alpha$.

Remarque :

On constate que plus l'intervalle de confiance est étroit, plus le risque augmente ; tandis que plus l'intervalle de confiance a une grande amplitude, plus le risque est faible. En pratique, il faudra trouver un compromis entre les deux. En pratique, un risque de 0,05 est acceptable... on cherchera donc souvent un intervalle de confiance de niveau de confiance $0,95(95\%)$.

Méthode :

Recherche d'un intervalle de confiance exact pour $\mathrm{E}(X)$ avec l'inégalité de Bienaymé-Tchebychev.

Ici X admet une espérance m inconnue et une variance σ^2. On considère T_n un estimateur **sans biais** de m, c'est à dire tel que $\mathbb{E}\left(T_n\right) = \boldsymbol{m}$. Pour obtenir un intervalle de confiance :

1. Inégalité de Bienaymé-Tchebychev : $\forall \varepsilon > 0, \mathbb{P}\left(|T_n - m| \geq \varepsilon\right) \leq \dfrac{\mathbb{V}\left(T_n\right)}{\varepsilon^2}$.
2. Si $\mathbb{V}\left(T_n\right)$ dépend de m, on majore $\mathbb{V}\left(T_n\right)$ par v_n, indépendant de m (sinon, on prend $v_n = \mathbb{V}\left(T_n\right)$) et ainsi :
$$\forall \varepsilon > 0, \mathbb{P}\left(|T_n - m| \geq \varepsilon\right) \leq \frac{v_n}{\varepsilon^2}$$
3. Par passage à l'évènement contraire, on obtient :
$$\forall \varepsilon > 0, \mathbb{P}\left(|T_n - m| < \varepsilon\right) \geq 1-\frac{v_n}{\varepsilon^2} \quad \text{autrement dit} \quad \forall \varepsilon > 0, \mathbb{P}\left([T_n - \varepsilon < m < T_n + \varepsilon]\right) \geq 1-\frac{v_n}{\varepsilon^2}$$
4. Et puisque $[T_n - \varepsilon < m < T_n + \varepsilon] \subset [T_n - \varepsilon \leq m \leq T_n + \varepsilon]$, on obtient
$$\forall \varepsilon > 0, \mathbb{P}\left([m \in [T_n - \varepsilon; T_n + \varepsilon]]\right) \geq 1-\frac{v_n}{\varepsilon^2}$$
Reste à choisir ε tel que, pour le α donné, on ait $\dfrac{v_n}{\varepsilon^2} = \alpha \ldots$ et ainsi :
$$\mathbb{P}\left([m \in [T_n - \varepsilon; T_n + \varepsilon]]\right) \geq 1-\alpha$$
L'intervalle de confiance proposé est centré en T_n, d'amplitude 2ε.

Remarques :

R1 – En se fixant un risque $\alpha = \dfrac{1}{4n\varepsilon^2}$, plus ε diminue et plus n doit être grand pour conserver le même risque.

R2 – En prenant $\alpha = \dfrac{1}{4n\varepsilon^2}$, on voit aussi que, à n fixé, plus ε est petit et plus α est grand. et plus ε est grand, plus α est petit.

Exercice 7

On considère une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu ainsi qu'un n échantillon $(X_1, \dots, X_n)$ de X. On note $\overline{X_n}$ la moyenne empirique de cet échantillon.

1. Montrer que : $\forall \varepsilon > 0, \mathbb{P}\left(\left[\overline{X_n} - \varepsilon \leq p \leq \overline{X_n} + \varepsilon\right]\right) \geq 1 - \dfrac{1}{4n\varepsilon^2}$.
2. En déduire que $\left[\overline{X_n} - \sqrt{\dfrac{5}{n}}, \overline{X_n} + \sqrt{\dfrac{5}{n}}\right]$ est un intervalle de confiance de p au niveau de confiance 95%.
3. Au second tour d'une élection présidentielle, les citoyens ont le choix entre deux candidats A et B. Un institut de sondage réalise un sondage auprès de 1000 personnes, dont 570 affirment vouloir voter pour le candidat A. Peut-on affirmer, avec un risque d'erreur de 5% que la candidat A sera élu ?

III. 2 Intervalle de confiance asymptotique

Définition 3.2

Soit X une variable aléatoire suivant une loi dépendant d'un certain paramètre θ inconnu. Soient $(U_n)_{n\in\mathbb{N}^*}$ et $(V_n)_{n\in\mathbb{N}^*}$ deux suites d'estimateurs de θ telles que : $\forall n \in \mathbb{N}^*, \mathbb{P}_\theta([U_n \leq V_n]) = 1$.
Soit $\alpha \in]0;1[$. On dit que $[U_n, V_n]$ est **un intervalle de confiance asymptotique de θ au niveau de confiance $1-\alpha$ (ou au risque α)**, lorsque

$$\lim_{n\to+\infty} \mathbb{P}_\theta([\theta \in [U_n, V_n]]) \geq 1 - \alpha$$

Remarque :

Les intervalles de confiance asymptotiques sont toujours déduits d'une convergence en loi. En particulier, le théorème central limite est très utile lorsque l'estimateur T_n est la moyenne empirique.

Méthode :

Recherche d'un intervalle de confiance asymptotique pour $\mathbb{E}(X)$ avec le TCL.

Ici X admet une espérance m inconnue et une variance σ^2 non nulle. On travaille avec l'estimateur $\overline{X_n}$ de $\boldsymbol{m}$ (estimateur sans biais). Pour obtenir un intervalle de confiance asymptotique :

1. En posant $\overline{X_n}^* = \sqrt{n}\dfrac{\overline{X_n} - m}{\sigma}$, on a, par le TCL : $\bar{X}_n^* \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} Z$, où $Z \hookrightarrow \mathcal{N}(0;1)$
2. Ainsi : $\forall x \geq 0, \quad \lim\limits_{n\to+\infty} \mathbb{P}\left(\left[-x \leq \bar{X}_n{}^* \leq x\right]\right) = \mathbb{P}([-x \leq Z \leq x]) = 2\Phi(x) - 1$.
3. Or $\forall x \geq 0, \mathbb{P}\left(\left[-x \leq \overline{X_n}^* \leq x\right]\right) = \ldots = \mathbb{P}\left(\left[\overline{X_n} - \dfrac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \dfrac{\sigma x}{\sqrt{n}}\right]\right)$

   D'où : $\lim\limits_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \dfrac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \dfrac{\sigma x}{\sqrt{n}}\right]\right) = 2\Phi(x) - 1$.
4. Si σ dépend de m, on majore σ par s, indépendant de m (sinon, on prend $s = \sigma$) et ainsi :

   $$\left[\overline{X_n} - \frac{\sigma x}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{\sigma x}{\sqrt{n}}\right] \subset \left[\overline{X_n} - \frac{sx}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{sx}{\sqrt{n}}\right]$$

   D'où, par croissance de P

   $$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X_n} - \frac{sx}{\sqrt{n}} \leq m \leq \overline{X_n} + \frac{sx}{\sqrt{n}}\right]\right) \geq 2\Phi(x) - 1$$
5. Reste à choisir x tel que, pour le α donné, on ait $2\Phi(x) - 1 = 1 - \alpha$. Mais :

$$2\Phi(x)-1=1-\alpha \iff \Phi(x)=1-\frac{\alpha}{2}$$
$$\iff \quad x=\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \qquad \Phi \text{ est bijective de } \mathbb{R} \text{ dans }]0,1[, \text{ et } 1-\frac{\alpha}{2}\in]0,1[$$

En notant $t_\alpha=\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$, on obtient :

$$\lim_{n\to+\infty}\mathbb{P}\left(\left[m\in\left[\overline{X_n}-t_\alpha\frac{s}{\sqrt{n}},\overline{X_n}+t_\alpha\frac{s}{\sqrt{n}}\right]\right]\right)\geq 2\Phi\left(t_\alpha\right)-1=1-\alpha$$

L'intervalle de confiance asymptotique proposé est centré en $\bar{X}_n$, d'amplitude $2t_\alpha\frac{s}{\sqrt{n}}$

Remarques :

R1 – L'expression des bornes de l'intervalle de confiance ne doit pas faire apparaître m

R2 – Avec $\alpha=0,05$, on a $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\simeq 1,96$.

R3 – De façon générale, σ est inconnu et on peut alors utiliser un estimateur de l'écart-type dans l'intervalle de confiance asymptotique..

Exercice 8

On considère une variable aléatoire suivant une loi de Bernoulli de paramètre p inconnu ainsi qu'un n échantillon $(X_1,\dots,X_n)$ de X. On note $\overline{X_n}$ la moyenne empirique de cet échantillon.

1. Justifier que :
$$\forall x\geq 0,\ \lim_{n\to+\infty}\mathbb{P}\left(\left[-x\leq\sqrt{n}\frac{\overline{X_n}-p}{\sqrt{p(1-p)}}\leq x\right]\right)=\Phi(x)-1$$
où Φ est la fonction de répartition d'une variable aléatoire suivant la loi $\mathcal{N}(0;1)$.
2. Soit $\alpha\in]0\,;\,1[$. Démontrer l'existence d'un unique réel t_α tel que $\Phi\left(t_\alpha\right)=1-\frac{\alpha}{2}$. On donne : $t_{0,05}\simeq 1,96$
3. En déduire que $\left[\overline{X_n}-\sqrt{\frac{1}{n}},\overline{X_n}+\sqrt{\frac{1}{n}}\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance 95%.
4. Quelle taille d'échantillon devons-nous avoir pour que l'intervalle de confiance asymptotique donné ait une amplitude inférieure ou égale à 0,02 ?
5. Comparons cet intervalle avec celui obtenu dans l'exemple 9 .

Paramètre d'une Bernoulli. Comparaison des intervalles de confiance

On considère un n-échantillon $(X_1,\dots,X_n)$ d'une loi de Bernoulli $X\hookrightarrow\mathcal{B}(p)$. On cherche un intervalle de confiance pour p au risque α. On sait que $T_n=\bar{X}_n$ est un estimateur non biaisé de p.

- D'après ce que l'on a vu, l'inégalité de Bienaymé-Tchébychev fournit un intervalle de confiance
$$P\left(p\in\left[T_n-\sqrt{\frac{1}{4\alpha n}};T_n+\sqrt{\frac{1}{4\alpha n}}\right]\right)\geq 1-\alpha$$
- On peut aussi procéder avec le théorème central limite. Ce dernier permet d'affirmer que
$$T_n^*=\sqrt{\frac{n}{p(1-p)}}\left(T_n-p\right)\xrightarrow{\mathcal{L}}Z,\quad Z\hookrightarrow\mathcal{N}(0;1)$$

On sait que $\sigma^2=p(1-p)\leq 1/4$, et donc d'après ce qui précède on obtient un autre intervalle de confiance, celui-ci asymptotique, au risque α

$$\lim_{n\to+\infty}P\left(p\in\left[T_n-\frac{1}{2\sqrt{n}}t_\alpha;T_n+\frac{1}{2\sqrt{n}}t_\alpha\right]\right)\geq 1-\alpha$$

Le premier intervalle à l'avantage d'être exact, non asymptotique. Cependant le deuxième, pour peu que n soit assez grand, a une longueur plus faible et donne des estimations plus précises de p. On se demande alors *à partir de quelle valeur de n est-il raisonnable d'échanger le caractère exact de l'intervalle de confiance contre un intervalle plus réduit ?* On considère qu'en pratique, on doit au moins avoir $n > 20$.

Simulation informatique :

Nous allons comparer les intervalles de confiance du paramètre d'une loi de Bernoulli obtenus avec le TCL et l'inégalité de Bienaymé-Tchebychev.

1. Pour $n = 100$, ces deux intervalles s'écrivent :

$$\left[\overline{X_{100}} - \frac{1}{20\sqrt{\alpha}}, \overline{X_{100}} + \frac{1}{20\sqrt{\alpha}}\right] \text{ et } \left[\overline{X_{100}} - \frac{1}{20}\Phi^{-1}\left(1-\frac{\alpha}{2}\right), \overline{X_{100}} + \frac{1}{20}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

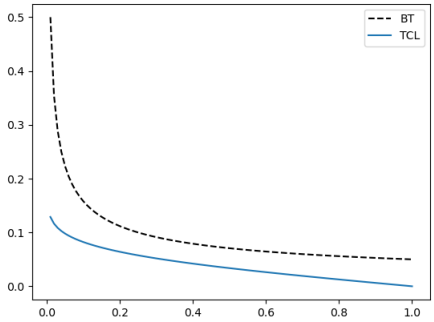
Les demi-longueurs de ces intervalles correspondent respectivement à

$$\frac{1}{20\sqrt{\alpha}} \qquad \text{et} \qquad \frac{1}{20}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

Nous allons faire varier le risque α et regarder comment évolue ces demi-longueurs :

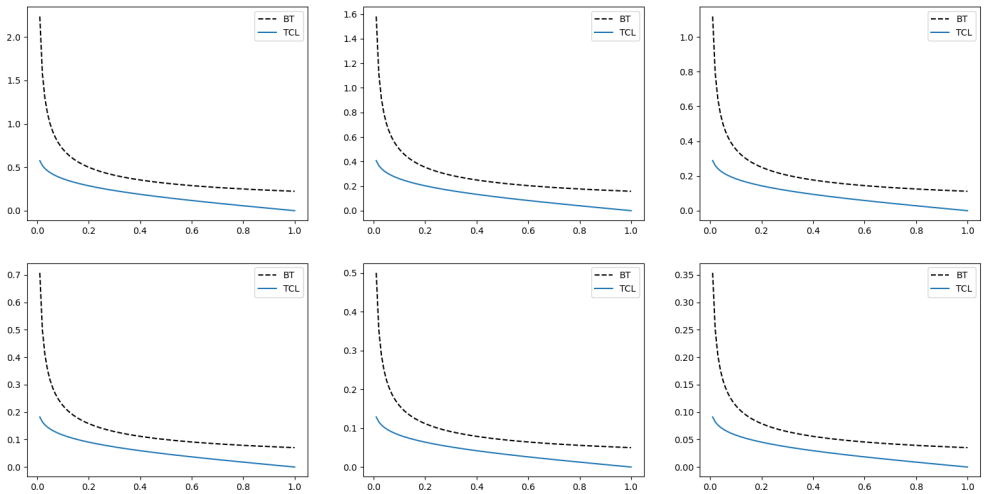
```
52 p=rd.rand()
53 alpha=np.linspace(0.01,1,100)
54 plt.plot(alpha,1/(20*np.sqrt(alpha)),'k--',label="BT")
55 plt.plot(alpha,sp.ndtri(1-alpha/2)/20, label="TCL")
56 plt.legend()
57 plt.show()
```

Python affiche alors



On constate que la première courbe est toujours au dessus de la seconde, ce qui signifie que pour toute valeur de α, l'intervalle de confiance obtenu par l'inégalité de B-T est toujours plus large que celui obtenu par le TCL. Il est donc moins bon.

A niveau de confiance égal, l'intervalle de confiance obtenu par le TCL est toujours meilleur que celui obtenu par l'inégalité de B-T. Ceci se vérifie pour différentes tailles d'échantillons. On a fait varier n dans la liste `[5,10,20,50,100,200]` :



et la conclusion semble se confirmer.

2. Nous allons à présent simuler 1000 échantillons de taille $n = 500$, construire les 1000 intervalles de confiance obtenus avec l'inégalité de B-T, puis avec le TCL et déterminer la proportion d'intervalles contenant p. Et nous allons répéter cela 10 fois.

   Avec $n = 500$ et $\alpha = 0.05$, l'intervalle de confiance donné par l'inégalité de B-T est :

$$\left[\overline{X_{500}} - \frac{1}{10}, \overline{X_{500}} + \frac{1}{10}\right] \text{ et celui donnée par le TCL est : } \left[\overline{X_{500}} - \frac{t_{0.05}}{2\sqrt{500}}, \overline{X_{500}} + \frac{t_{0.05}}{2\sqrt{500}}\right]$$

   On tape dans Python les commandes suivantes

```
t=sp.ndtri(1-0.05/2) #Bijection réciproque de la fonction de répartition d'une loi n
c=t/(2*np.sqrt(500)) #Demi-longueur pour le TCL

Xbar=np.zeros(1000) # Création d'un vecteur de taille 100
for j in range(10):
    p=rd.rand()
    for i in range(1000):
        Xbar[i]=np.mean(rd.binomial(1, p,500))
    BT=np.mean((Xbar-1/10 <=p)*(Xbar+0.1>=p))
    TCL=np.mean((Xbar-c<=p)*(Xbar+c>=p))
    print('p=',p,"Fréquende dans IC avec BT",BT,"Fréquence dans IC avec TCL",TCL)
```

Remarque :

> Les commandes `(Xbar-1/10 <=p)*(Xbar+0.1>=p)` et `(Xbar-c<=p)*(Xbar+c>=p)` renvoient des vecteurs booléens, avec un 1 si p est dans l'intervalle, 0 sinon. Ainsi `BT` et `TCL` donnent une valeur approchée de la probabilité que l'intervalle de confiance en question contienne effectivement le paramètre que l'on voulait estimer.

Python affiche alors :

```
>>> (executing cell "Comparaisons d'in..." (line 93 of "Estimation.py"))
p= 0.6405686910287822 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.953
p= 0.6796650734176325 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.969
p= 0.7338280404762813 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.97
p= 0.2708474380123379 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.974
p= 0.3436730587247212 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.961
p= 0.14424280837525882 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.993
p= 0.19911207071945325 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.981
p= 0.8667383420125998 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.996
p= 0.6137404467618962 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.952
p= 0.6331969568529442 Fréquende dans IC avec BT 1.0 Fréquence dans IC avec TCL 0.954
```

On constate qu'avec l'intervalle de confiance pour Bienaymé-Tchebytchev, on a 100% de succès, ce n'est pas étonnant car l'intervalle obtenu est trop grand, d'amplitude égale à 0.2 . Il n'est donc pas très intéressant. L'intervalle obtenu avec le TCL est lui d'amplitude égale 0.088 , il est donc plus difficile d'y trouver la valeur de p. La réponse donnée par Python est toujours de l'ordre de 95% .