Table des matières

Dans tout le chapitre, toutes les variables aléatoires considérées (discrètes ou à densité) sont définies sur un même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.

1 Introduction

Exemple 1. Si on cherche à calculer la taille moyenne d'un homme adulte en France, il est impossible (pour des questions de rentabilité) de déterminer la taille de tous les français adultes et d'en faire la moyenne. Pour donner une valeur approchée de cette moyenne, on prend un échantillon d'hommes (on sonde par exemple 10000 individus), on détermine leur taille, puis on fait la moyenne des tailles obtenues. Avec un échantillon assez grand, on considère que l'on a obtenu une valeur approchée, ou **une estimation (ponctuelle)**, de la taille moyenne d'un homme adulte en France.

Exemple 2. On dispose d'une pièce de monnaie, et on se demande si la pièce est équilibrée. On lance 100 fois la pièce et on obtient, au cours de ces 100 lancers, 40 `Pile`. On considère alors que $\frac{40}{100}$ est une valeur approchée, ou une **estimation (ponctuelle)**, de la probabilité p de l'événement « la pièce tombe sur `Pile` ». Au vu de cette observation, on peut penser que la pièce est truquée.

Le problème de l'estimation. On cherche à étudier une caractéristique θ (ou $g(\theta)$, avec g une fonction) d'une population de taille N. Par exemple, on se demande quelle est la taille moyenne θ d'un homme français, si une pièce de monnaie est truquée, quel candidat va remporter les élections, quelle est la température moyenne à Paris au mois de mars... La plupart du temps, nous ne pouvons pas déterminer la valeur du paramètre θ, pour des raisons de temps et de rentabilité (on ne peut pas recenser la taille de chaque français...) ou pour cause du caractère aléatoire du phénomène étudié (température moyenne à Paris). À l'aide d'une étude statistique (réalisation d'un sondage), nous allons alors chercher à estimer le paramètre θ (ou $g(\theta)$) qui nous intéresse. On constitue alors un échantillon représentatif de la population initiale, afin de pouvoir effectivement observer les caractéristiques nécessaires à l'étude de θ (ou $g(\theta)$) sur les individus de l'échantillon. En général, la taille n de cet échantillon est très petite devant N.

Pour répondre au problème de l'estimation du paramètre θ (ou $g(\theta)$), nous modélisons le phénomène observé par une expérience aléatoire et des variables aléatoires qui lui sont associées. Par exemple :

- Problème 1 :
  On se demande quelle est la taille moyenne d'un homme français. On considère que la taille d'un homme français est une variable aléatoire X qui suit la loi normale de paramètres m et σ^2. On cherche donc à estimer m (c'est-à-dire à déterminer une valeur approchée de m).
- Problème 2 :
  On se demande si une pièce de monnaie est truquée. On considère une variable aléatoire X qui suit la loi de Bernoulli paramètre p : X prend la valeur 1 si la pièce amène `Pile`, et X prend la valeur 0 si la pièce amène `Face`. Ici, on cherche à estimer p (et on se demande si p est égal à $\frac{1}{2}$).
- Problème 3 :
  On se demande si le candidat A va remporter les élections. On cherche donc la proportion p des électeurs qui votent pour A. On considère une variable aléatoire X de loi de Bernoulli de paramètre p : X prend la valeur 1 si un électeur choisi vote pour A et X prend la valeur 0 sinon. Ici, on cherche à estimer p (et à savoir si $p > 0,5$).
- Problème 4 :
  On se demande quel est le temps d'attente moyen d'un client à la poste (faut-il ouvrir de nouveaux guichets ?). On considère une variable aléatoire X égale au temps d'attente d'un client à la poste, et on suppose que X suit la loi exponentielle de paramètre λ. On cherche donc à estimer $g(\lambda) = \frac{1}{\lambda}$ (car $\frac{1}{\lambda} = \mathbb{E}(X)$ est le temps moyen d'attente du client).

Bilan. On s'intéresse à une variable aléatoire X dont la loi dépend d'un paramètre θ (réel ou vectoriel) qui est inconnu ($\theta = m$ dans le problème 1, $\theta = p$ dans les problèmes 2 et 3 et $\theta = \lambda$ dans le problème 4), et on cherche à estimer θ ou $g(\theta)$ (par exemple $g(\lambda) = \frac{1}{\lambda}$ dans le problème 4). Par exemple, la variable X suit la loi :

- ⋆ $\mathcal{B}(\theta)$, avec $\theta \in]0,1[$ inconnu ;
- ⋆ $\mathcal{U}([-\theta,\theta])$, avec $\theta > 0$ inconnu ;
- ⋆ $\mathcal{P}(\theta)$, avec $\theta > 0$ inconnu ;
- ⋆ $\mathcal{N}(\theta,1)$, avec $\theta > 0$ inconnu ;
- ⋆ $\mathcal{N}\left(0,\theta^2\right)$, avec $\theta > 0$ inconnu ;
- ⋆ $\mathcal{N}\left(m,\sigma^2\right)$, avec m et $\sigma > 0$ inconnus (ici $\theta = (m,\sigma)$ est un paramètre vectoriel)...

La modélisation pour l'estimation. On considère une expérience aléatoire $\mathscr{E}$ et une variable aléatoire X qui lui est liée. On suppose que la loi de X dépend d'un paramètre réel θ inconnu (ou parfois d'un paramètre vectoriel θ) : on ne connaît pas la loi de X, mais on sait que la loi de X appartient à une famille de loi μ_θ dépendant d'un paramètre réel θ (ou d'un paramètre vectoriel θ), avec $\theta \in \Theta$ où Θ est un sous-ensemble de $\mathbb{R}$ (ou éventuellement de $\mathbb{R}^2$). Par exemple, pour les problèmes 2 et 3, on a $\mu_\theta = \mathcal{B}(\theta)$, et $\Theta =]0,1[$.

On cherche à estimer la valeur de θ (ou de $g(\theta)$).

Lorsque l'on réalise une fois l'expérience $\mathscr{E}$, la valeur que prend la variable aléatoire X, notée x, s'appelle une **réalisation** de X. La seule réalisation x de X ne permet pas de donner une valeur approchée de θ (ou de $g(\theta)$) : si l'on cherche à savoir si le candidat A va remporter les élections, on ne peut pas réaliser un sondage auprès d'un unique électeur !

Ainsi, pour obtenir une estimation de θ (ou de $g(\theta)$), on répète n fois la même expérience $\mathscr{E}$ dans des conditions identiques et indépendantes, et on note $(x_1, x_2, \dots, x_n)$ les n réalisations observées de la variable aléatoire X au cours des ces n expériences. On dit alors que $(x_1, x_2, \dots, x_n)$ est un ***n*-échantillon de données**, ou un ***n*-échantillon de réalisations** de la variable aléatoire X. De plus, en répétant n fois la même expérience $\mathscr{E}$ dans des conditions identiques et indépendantes, on définit n variables aléatoires $X_1, X_2, \dots, X_n$ mutuellement indépendantes et de même loi que X : pour tout $k \in [\![1,n]\!]$, X_k est la variable aléatoire associée à la k^{e} expérience $\mathscr{E}$ effectuée, et x_k est la réalisation de la variable aléatoire X_k.
On dispose alors de n variables aléatoires $X_1, X_2, \dots, X_n$ indépendantes et de même loi que X (dépendant d'un paramètre θ inconnu, que l'on cherche à estimer, au vu des réalisations des variables aléatoires $X_1, X_2, \dots, X_n$). On dit alors que $(X_1, X_2, \dots, X_n)$ est un ***n*-échantillon indépendant et identiquement distribué de la loi de *X***.

Comment estimer θ ? (reprise de l'exemple 2 et du problème 2) On cherche ici à estimer la probabilité p que la pièce amène `Pile`. On note, pour tout $i \in [\![1,100]\!]$, X_i la variable aléatoire de Bernoulli égale à 1 si l'on obtient `Pile` au i^{e} lancer : $(X_1, X_2, \dots, X_{100})$ est un 100-échantillon indépendant et identiquement distribué de la loi de Bernoulli $\mathcal{B}(p)$. On lance 100 fois la pièce et on obtient, au cours de ces 100 lancers, 40 `Pile` : on dispose d'un 100-échantillon de données $(x_1, x_2, \dots, x_{100})$ tel que parmi ces 100 réels 40 sont égaux à 1 et 60 sont égaux à 0.

On considère alors que $\hat{p} = \dfrac{40}{100} = \dfrac{1}{100}\displaystyle\sum_{k=1}^{100} x_k$ est une valeur approchée, ou une **estimation (ponctuelle)**, de la probabilité p. Ici, on a utilisé l'**estimateur** $T_{100} = \dfrac{1}{100}\displaystyle\sum_{k=1}^{100} X_k$ pour estimer p, et $\hat{p} = \dfrac{40}{100} = \dfrac{1}{100}\displaystyle\sum_{k=1}^{100} x_k$ est une **réalisation de l'estimateur** T_{100}, autrement dit une **estimation** de p. Notons que, d'après la loi faible des grands nombres, l'estimation $\hat{p}$ est une valeur approchée de p avec forte probabilité.

Bilan. Soit X une variable aléatoire dont la loi dépend d'un paramètre θ inconnu. On cherche à estimer la valeur de θ (ou de $g(\theta)$). On dispose d'un ***n*-échantillon $(X_1, X_2, \dots, X_n)$ de variables aléatoires indépendantes et de même loi que *X***.

L'**estimation ponctuelle** consiste à déterminer une valeur approchée de θ (ou de $g(\theta)$) lorsque l'on dispose d'une réalisation de chacune des variables aléatoires $X_1, X_2, \dots, X_n$, c'est-à-dire d'un ***n*-échantillon de données $(x_1, x_2, \dots, x_n)$**.
On va chercher une variable aléatoire T_n fonction des variables aléatoires $X_1, X_2, \dots, X_n$: la variable aléatoire $T_n = \varphi(X_1, X_2, \dots, X_n)$ est appelée un **estimateur** de θ. La valeur $\hat{\theta} = \varphi(x_1, x_2, \dots, x_n)$ est appelée une **estimation (ponctuelle)** de θ : $\hat{\theta}$ est une réalisation de l'estimateur T_n. Le but est donc de rechercher des estimateurs T_n de θ afin que l'estimation $\hat{\theta}$, valeur observée, soit la plus proche possible de la valeur théorique de θ. Nous allons dans ce cours donner des exemples d'estimateurs et décrire une méthode pour construire un estimateur de θ (appelé **estimateur du maximum de vraisemblance**). Les critères (tels que le **biais** d'un estimateur et la notion d'**estimateur convergent**) permettant de juger de la qualité d'un estimateur sont hors-programme.

Une méthode plus efficace pour estimer θ (ou $g(\theta)$) consiste à chercher un intervalle aléatoire $[U_n, V_n]$, où U_n et V_n sont deux estimateurs de θ (ou de $g(\theta)$), contenant le paramètre θ (ou $g(\theta)$) avec une très forte probabilité. On dit dans ce cas que l'on construit un **intervalle de confiance** de θ (ou de $g(\theta)$). Nous verrons dans ce cours comment construire des intervalles de confiance à l'aide de l'inégalité de Bienaymé-Tchebychev ou du théorème central limite.

2 Estimation ponctuelle

Notations. X désigne une variable aléatoire réelle dont la loi dépend d'un paramètre θ réel (ou vectoriel) inconnu. On suppose que θ appartient à Θ un sous-ensemble de $\mathbb{R}$ (ou de $\mathbb{R}^2$). On note $(\mu_\theta)_{\theta\in\Theta}$ la famille de loi de probabilité à laquelle appartient la loi de X. Parfois, l'espace probabilisé sur lequel est défini la variable aléatoire X est noté $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$.

Dans tout le cours, g désigne une fonction définie sur Θ.

Exemple 3.

- ⋆ $\mu_\theta = \mathcal{B}(\theta)$, avec $\Theta =]0,1[$;
- ⋆ $\mu_\theta = \mathcal{U}([0,\theta])$, avec $\Theta = \mathbb{R}_+^*$;
- ⋆ $\mu_\theta = \mathcal{P}(\theta)$, avec $\Theta = \mathbb{R}_+^*$;
- ⋆ $\mu_\theta = \mathcal{N}(\theta, 1)$, avec $\Theta = \mathbb{R}$;
- ⋆ $\mu_\theta = \mathcal{N}\left(m, \sigma^2\right)$, avec $\theta = (m, \sigma^2)$ et $\Theta = \mathbb{R} \times \mathbb{R}_+^*$;

2.1 Notion de n-échantillon

Definition 1. Soit $n \in \mathbb{N}^*$. On appelle *n-échantillon* de variables aléatoires indépendantes et de même loi que X (ou *n-échantillon* indépendant, identiquement distribué de la loi μ_θ) toute famille de n variables aléatoires $X_1, \dots, X_n$ telles que :

1. les variables $X_1, X_2, \dots, X_n$ sont mutuellement indépendantes ;
2. pour tout entier $i \in [\![1, n]\!]$, X_i suit la même loi que X (i.e. $X_i \hookrightarrow \mu_\theta$).

Remarque 1. La loi de X dépendant du paramètre θ, on notera parfois $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$ l'espace probabilisé sur lequel est défini la variable aléatoire X et le n-échantillon $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et de même loi que X. On dit dans ce cas que : pour tout $\theta \in \Theta$, les variables aléatoires $X_1, X_2, \dots, X_n$ sont $\mathbb{P}_\theta$-indépendantes.
De même, lorsque X admet une espérance (resp. une variance), celle-ci sera notée $\mathbb{E}_\theta(X)$ (resp. $\mathbb{V}_\theta(X)$) et $\mathbb{E}_\theta(X_i)$ (resp. $\mathbb{V}_\theta(X_i)$) désigne l'espérance (resp. la variance) de X_i.
Ces notations sous-entendent que l'on travaille sous l'hypothèse que $X \hookrightarrow \mu_\theta$, avec $\theta \in \Theta$ fixé.

Definition 2. Soit $n \in \mathbb{N}^*$. Soit $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. On dit que $(x_1, \dots, x_n)$ est un *n-échantillon de données* (ou un *n-échantillon de réalisations* de la variable aléatoire X) lorsque, pour tout $i \in [\![1, n]\!]$, x_i est une réalisation de la variable aléatoire X_i, où $(X_1, \dots, X_n)$ est un n-échantillon de variables aléatoires indépendantes et de même loi que X.

Remarque 2. Un n-échantillon $(x_1, \dots, x_n)$ de données est obtenu en réalisant n expériences identiques et indépendantes.

Exemple 4. On se demande si une pièce est truquée. On note X la variable aléatoire de Bernoulli égale à 1 si le lancer donne `Pile`. Ainsi, X suit la loi de Bernoulli $\mathcal{B}(p)$, avec $p \in]0,1[$ un paramètre inconnu. On lance cinq fois la pièce et l'on obtient : `Pile-Pile-Face-Pile-Face`. Dans ce cas, $(1, 1, 0, 1, 0)$ est un 5-échantillon de données de réalisations de la variable aléatoire X.

2.2 Estimateur et estimation

Definition 3.

- Soit $(X_1, \dots, X_n)$ un n-échantillon de variables aléatoires indépendantes et de même loi que X.
  Si $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ est une application dont l'expression ne dépend pas de θ, alors on dit que la variable aléatoire réelle $T_n = \varphi(X_1, \dots, X_n)$ est un estimateur de $g(\theta)$.
- Plus généralement, une suite d'estimateurs de $g(\theta)$ est une suite de variables aléatoires réelles $(T_n)_{n\in\mathbb{N}^*}$ telle que, pour tout $n \in \mathbb{N}^*$, $T_n = \varphi_n(X_1, \dots, X_n)$ avec $\varphi_n : \mathbb{R}^n \longrightarrow \mathbb{R}$ ne dépendant pas de θ et $(X_1, \dots, X_n)$ un n-échantillon de variables aléatoires indépendantes et de même loi μ_θ.

Remarque 3.

Un estimateur T_n de $g(\theta)$ est une fonction de n variables aléatoires réelles indépendantes et de même loi μ_θ.

La loi de la variable aléatoire T_n dépend du paramètre θ, mais l'expression de T_n ne dépend pas de θ. Autrement dit, la fonction $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$ peut éventuellement dépendre de n, mais ne dépend pas du paramètre inconnu θ.
Parfois, par abus de langage, une suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ de $g(\theta)$ sera appelée plus simplement un estimateur de $g(\theta)$.

Definition 4. Soit $T_n = \varphi(X_1, \dots, X_n)$ un estimateur de $g(\theta)$, avec $\varphi : \mathbb{R}^n \longrightarrow \mathbb{R}$.
Si $(x_1, \dots, x_n)$ est un n-échantillon de réalisations de l'échantillon $(X_1, \dots, X_n)$, alors on dit que $t_n = \varphi(x_1, \dots, x_n)$ est une estimation (ponctuelle) de $g(\theta)$.

Remarque 4. Une estimation de $g(\theta)$ est une réalisation d'un estimateur de $g(\theta)$. Une estimation ponctuelle de $g(\theta)$ dépend des réalisations observées lors de n expériences aléatoires identiques et mutuellement indépendantes. Deux réalisations de l'estimateur T_n donnent deux estimations distinctes de $g(\theta)$.

2.3 Exemples d'estimateurs

Méthode de la moyenne empirique. On suppose ici que la variable aléatoire X admet une espérance $\mathbb{E}_\theta(X)$ et une variance $\mathbb{V}_\theta(X)$. Comme $X_1, \dots, X_n$ sont des variables aléatoires indépendantes, admettant la même espérance $\mathbb{E}_\theta(X)$ et la même variance, d'après la loi faible des grands nombres :

$$\forall \varepsilon > 0, \lim_{n \to +\infty} \mathbb{P}\left(\left[\left|\overline{X}_n - \mathbb{E}_\theta(X)\right| \geqslant \varepsilon\right]\right) = 0$$

Ainsi, lorsque n est grand, une réalisation de $\overline{X}_n$ est très proche de la valeur $\mathbb{E}_\theta(X)$ avec grande probabilité.

La *moyenne empirique* $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ est donc un estimateur « naturel » de $g(\theta) = \mathbb{E}_\theta(X)$.

Exercice 1 : Soit X une v.a.r. dont la loi dépend d'un paramètre θ. On suppose que X admet une variance et on note $\mathbb{E}(X) = m$ et $\mathbb{V}(X) = \sigma^2$. Soit $n \in \mathbb{N}^*$ et soit $(X_1, \dots, X_n)$ un n-échantillon de la v.a.r. X. Calculer $\mathbb{E}(\overline{X}_n)$ et $\mathbb{V}(\overline{X}_n)$.

Exercice 2 :

1. On suppose que X suit la loi $\mathcal{B}(p)$. Proposer un estimateur de p.
2. On suppose que X suit la loi $\mathcal{P}(\lambda)$. Proposer un estimateur de λ.
3. On suppose que X suit la loi $\mathcal{G}(p)$. Proposer un estimateur de p.
4. On suppose que X suit la loi $\mathcal{U}([0, \theta])$. Proposer deux estimateurs de θ.
5. On suppose que X suit la loi $\mathcal{N}(0, \sigma^2)$. Proposer un estimateur de σ^2.
6. On suppose que X suit la loi $\mathcal{N}(m, \sigma^2)$, avec $m \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+^*$ deux paramètres inconnus. Proposer un estimateur de m, puis un estimateur de σ^2.

2.4 L'estimateur du maximum de vraisemblance

2.4.1 Le cas discret

La méthode du *maximum de vraisemblance* permet de construire des estimateurs intéressants. Détaillons ici le principe de cette méthode.

Considérons que l'on dispose d'une observation $(x_1, \dots, x_n)$ d'un n-échantillon $(X_1, \dots, X_n)$ d'une loi discrète de paramètre θ et que l'on cherche à estimer θ. L'idée est alors de choisir comme estimateur

$$\hat{\theta}_n = \varphi(X_1, \dots, X_n)$$

une fonction du n-échantillon $(X_1, .., X_n)$ où l'expression de la fonction φ est choisie de telle sorte que $\theta^\star = \varphi(x_1, \dots, x_n)$ soit la valeur rendant maximale la probabilité de l'évènement

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

(On cherche quelle valeur du paramètre θ maximise la probabilité d'observer ce que l'on a observé.)

Par hypothèse d'indépendance sur les variables du n-échantillon, la probabilité de l'évènement ci-dessus vaut

$$\mathbb{P}_\theta\left(\bigcap_{i=1}^{n} [X_i = x_i]\right) = \prod_{i=1}^{n} \mathbb{P}_\theta([X_i = x_i])$$

ce qui justifie les définitions ci-dessous.

Definition 5. Soient $(X_1, \dots, X_n)$ un n-échantillon d'une loi discrète de paramètre $\theta \in \Theta$ qui est un paramètre qu'on cherche à estimer et $(x_1, \dots, x_n) \in X_1(\Omega)^n$ fixé. La fonction L_n définie sur Θ par

$$L_n : \theta \longmapsto \prod_{i=1}^{n} \mathbb{P}_\theta([X_i = x_i])$$

s'appelle la *vraisemblance*.

En notant $\theta^* = \varphi(x_1, \dots, x_n)$ la valeur où L_n est maximale (c'est à dire telle que, pour tout $\theta \in \Theta$, $L_n(\theta) \leq L_n(\theta^*)$, l'*estimateur du maximum de vraisemblance* est l'estimateur défini par

$$\hat{\theta}_n = \varphi(X_1, \dots, X_n).$$

Exercice 3 : *(Estimateur du maximum de vraisemblance pour la loi de Bernoulli $\mathcal{B}(\theta)$)*

Soient $(X_1, \dots, X_n)$ un n-échantillon de la loi $\mathcal{B}(\theta)$ et $(x_1, \dots x_n) \in \{0; 1\}^n$. Ici le paramètre à estimer est $\theta \in]0, 1[$. On pose :

$$s_n = \sum_{i=1}^{n} x_i$$

1. Montrer que $L_n(\theta) = \theta^{s_n}(1-\theta)^{n-s_n}$.
2. On pose $h_n(\theta) = \ln(L_n(\theta))$.
   (a) Montrer que, pour tout $\theta \in]0, 1[$, $h_n'(\theta) = \frac{s_n - n\theta}{\theta(1-\theta)}$.
   (b) Montrer que h_n admet un maximum en un unique point θ^* que l'on explicitera en fonction de $x_1, \dots, x_n$.
3. Montrer que L_n admet également un maximum en θ^*.
4. Expliciter l'estimateur du maximum de vraisemblance. Quel estimateur reconnaît-on ?

Exercice 4 : *(Estimateur du maximum de vraisemblance pour la loi de Poisson $\mathcal{P}(\lambda)$).*

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
$$s_n = \sum_{i=1}^{n} x_i, \qquad p_n = \prod_{i=1}^{n} (x_i!),$$
exprimer la fonction de vraisemblance L_n de la loi de Poisson, définie sur $\mathbb{R}_+^*$.
2. Calculer $L_n'(\lambda)$ pour tout $\lambda > 0$ et vérifier que L_n est maximale en
$$\theta^* = \frac{s_n}{n}$$
3. Expliciter l'estimateur du maximum de vraisemblance. Quel estimateur reconnaît-on ?

Remarque 5. Dans les deux exemples précédents, c'est la *moyenne empirique* qui apparait être l'estimateur du maximum de vraisemblance, ce qui justifie aussi que ce soit un estimateur *naturel* à introduire (notamment dans notre problème des tanks). Cependant, ce n'est pas toujours le cas : l'estimateur du maximum de vraisemblance est parfois différent.

Exercice 5 : *(Estimateur du maximum de vraisemblance pour la loi géométrique $\mathcal{G}(p)$).*

Montrer que l'estimateur du maximum de vraisemblance de la loi $\mathcal{G}(p)$ est donné pour un n-échantillon $(X_1, \dots, X_n)$ par la formule :

$$\hat{\theta}_n = \frac{n}{X_1 + \dots + X_n}$$

2.4.2 Le cas à densité

On peut aussi définir la vraisemblance d'une loi à densité. En notant f_θ la densité d'une variable aléatoire X de loi de paramètre inconnue θ, et $(x_1, \dots, x_n) \in X(\Omega)^n$, il s'agit de la fonction

$$L_n : \theta \longmapsto \prod_{i=1}^{n} f_\theta(x_i)$$

L'estimateur du maximum de vraisemblance est alors défini de la même manière que précédemment.

Exercice 6 : On considère un n-échantillon de la loi $\mathcal{U}([0,\theta])$ et on cherche à estimer θ. Soit $(x_1, \dots, x_n) \in (\mathbb{R}_+^*)^n$. On note f_θ la densité usuelle de la loi $\mathcal{U}([0,\theta])$. On introduit la fonction de vraisemblance, définie sur $\mathbb{R}_+$ par

$$\forall \theta > 0,\ L_n(\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

1. Montrer que, pour tout $\theta > 0$, on a $L_n(\theta) = \begin{cases} \theta^{-n} & \text{si } \theta \geq \max(x_1, \dots, x_n) \\ 0 & \text{sinon} \end{cases}$.
2. Tracer le graphe de L_n sur $]0, +\infty[$.
3. En déduire que l'estimateur du maximum de vraisemblance pour la loi $\mathcal{U}([0,\theta])$ est donné par :

$$\hat{\theta}_n = \max(X_1, \dots, X_n)$$

Exercice 7 : `(extrait de EDHEC 2012)`

Soit $Y \hookrightarrow \mathcal{E}(\lambda)$ de densité f_Y. On suppose, dans la suite, que le paramètre λ est inconnu et on souhaite l'estimer en utilisant la loi de Y. On désigne par n un entier naturel supérieur ou égal à 2 et on considère n variables aléatoires $Y_1, \dots, Y_n$, supposées définies sur $(\Omega, \mathscr{A}, \mathbb{P})$. On suppose qu'elles sont indépendantes et de même loi que Y.

1. On considère des réels $x_1, \dots, x_n$ strictement positifs, ainsi que la fonction L, à valeurs dans $\mathbb{R}$, définie sur $]0, +\infty[$ par : $\forall \lambda \in]0, +\infty[$, $L(\lambda) = \prod_{k=1}^{n} f_Y(x_k)$.
   (a) Exprimer $L(\lambda)$, puis $\ln\big(L(\lambda)\big)$ en fonction de λ, $x_1, \dots, x_n$.
   (b) On considère la fonction φ, définie pour tout réel λ de $]0, +\infty[$ par : $\varphi(\lambda) = n \ln(\lambda) - \lambda \sum_{k=1}^{n} x_k$.
   Montrer que la fonction φ admet un maximum, atteint en un seul réel que l'on notera z et que l'on exprimera en fonction de $x_1, \dots, x_n$.
   Que peut-on dire de z pour la fonction L ?
2. On pose dorénavant, toujours avec n supérieur ou égal à 2 , $Z_n = \dfrac{n}{\sum_{k=1}^{n} Y_k}$.
   On admet que Z_n est une variable aléatoire définie, elle aussi, sur l'espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$.
   La suite $(Z_n)_{n\geq 2}$ est appelée estimateur du maximum de vraisemblance pour λ.
   (a) Pour tout $n \in \mathbb{N}^*$, on définit la variable aléatoire S_n par : $S_n = \sum_{k=1}^{n} Y_k$.
   On admet le résultat suivant :
   Soient X et Y deux variables aléatoires à densité indépendantes définies sur le même espace probabilisé, de densités respectives f_X et f_Y telles que f_X et f_Y soient bornées.
   Alors la variable aléatoire $X + Y$ est une variable aléatoire à densité et une densité de $X + Y$ est donnée par la fonction h définie sur $\mathbb{R}$ par :

$$h : x \mapsto \int_{-\infty}^{+\infty} f_X(t)\, f_Y(x - t)\, \mathrm{d}t$$

   En utilisant la propriété admise, montrer que, pour tout $n \in \mathbb{N}^*$, la variable aléatoire S_n est une variable aléatoire à densité et admet pour densité la fonction f_n définie par :

$$f_n : t \mapsto \begin{cases} 0 & \text{si } t < 0 \\ \dfrac{\lambda^n}{(n-1)!}\, t^{n-1} \mathrm{e}^{-\lambda t} & \text{si } t \geq 0 \end{cases}$$

(b) Soit $n \geq 2$. En remarquant que $\int_0^{+\infty} f_{n-1}(t)\,\mathrm{d}t = 1$, montrer que Z_n possède une espérance et que $\mathbb{E}(Z_n) = \dfrac{n}{n-1}\,\lambda$.

(c) Déterminer un estimateur Z_n' de λ, fonction simple de Z_n, qui soit un estimateur sans biais de λ (*i.e.* $\mathbb{E}(Z_n') = \lambda$).

3 Estimation par intervalle de confiance (exact ou asymptotique)

À chaque estimation (observation d'un estimateur), correspond une valeur approchée, de précision non spécifiée, du paramètre θ. On peut vouloir préciser l'erreur commise (ainsi que le risque d'erreur), c'est à dire déterminer un **intervalle**, plutôt qu'une unique valeur, contenant θ avec un haut niveau de confiance. C'est l'estimation par intervalle de confiance.

3.1 Définitions

Dans cette sous-partie :

- $(X_1, \dots, X_n)$ est un n-échantillon de X dont la loi dépend d'un paramètre θ
- pour tout $n \in \mathbb{N}^*$, $U_n = \varphi_n(X_1, \dots, X_n)$ et $V_n = \psi_n(X_1, \dots, X_n)$ sont des estimateurs de $g(\theta)$ tels que $\mathbb{P}_\theta\big([U_n \leq V_n]\big) = 1$

Definition 6 (Intervalle de confiance (exact))**.** Soit $\alpha \in]0,1[$.

- On dit que $[U_n, V_n]$ est un *intervalle de confiance de $g(\theta)$ au niveau de confiance* $1-\alpha$ si :
$$\mathbb{P}_\theta\Big([U_n \leq g(\theta) \leq V_n]\Big) \geq 1-\alpha$$
Le réel α est appelé le *niveau de risque de l'intervalle*, ou plus simplement le *risque*.
- Soit $\omega \in \Omega$ une issue. L'intervalle $[u_n, v_n] = [U_n(\omega), V_n(\omega)]$ est une réalisation de l'intervalle de confiance $[U_n, V_n]$, aussi appelé *intervalle de confiance observé*.

Remarque 6. On évitera de parler de probabilité pour l'intervalle de confiance observé, et on préférera le mot « risque » ou le mot « chance ». Par exemple, si $\alpha = 0,05$:

- Dire « il y a moins de 5% de risque que $g(\theta)$ ne soit pas dans l'intervalle de confiance observé $[u_n, v_n]$ » est correct.
- Dire « la probabilité que $g(\theta)$ soit dans $[u_n, v_n]$ est supérieure à $0,95$ » ou écrire « $\mathbb{P}_\theta\Big([u_n \leq g(\theta) \leq v_n]\Big) \geq 0,95$ » est maladroit.

En effet, l'événement $[u_n \leq g(\theta) \leq v_n]$ ne dépend d'aucune variable aléatoire, il est donc soit certain, soit impossible et on a soit $\mathbb{P}_\theta\Big([u_n \leq g(\theta) \leq v_n]\Big) = 1$, soit $\mathbb{P}_\theta\Big([u_n \leq g(\theta) \leq v_n]\Big) = 0$.

Les intervalles de confiance (exacts) seront, en pratique, toujours déduits de l'inégalité de Bienaymé-Tchebychev.

Definition 7 (Intervalle de confiance asymptotique)**.**

Soit $\alpha \in]0,1[$. On appelle *intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance* $1-\alpha$ toute suite $([U_n, V_n])_{n\in\mathbb{N}^*}$ d'intervalles aléatoires vérifiant : il existe une suite de réels $(\alpha_n)_{n\in\mathbb{N}^*}$ à valeurs dans $[0,1]$, de limite α, telle que :
$$\forall n \in \mathbb{N}^*,\ \mathbb{P}_\theta\Big([U_n \leq g(\theta) \leq V_n]\Big) \geq 1-\alpha_n$$

Remarque 7. On pourra dire que $[U_n, V_n]$ est un intervalle de confiance asymptotique de $g(\theta)$ au niveau de confiance $1-\alpha$ même s'il s'agit d'un abus de langage.

Les intervalles de confiance asymptotiques sont toujours déduits d'une convergence en loi. En particulier, le théorème central limite est très utile lorsque les estimateurs sont construits à partir de la moyenne empirique.

3.2 L'exemple fondamental du sondage : estimation par intervalle de confiance du paramètre d'une loi de Bernoulli

Soit $X \hookrightarrow \mathcal{B}(p)$. On suppose que le paramètre $p \in]0,1[$ est inconnu. Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes de même loi que X. On note, pour tout $n \in \mathbb{N}^*$, $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$.

Rappelons que $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X) = p$ et $\mathbb{V}(\overline{X}_n) = \dfrac{\mathbb{V}(X)}{n} = \dfrac{p(1-p)}{n}$.

3.2.1 Construction d'un intervalle de confiance (exact) via l'inégalité de Bienaymé-Tchebychev

Soit $\alpha \in]0,1[$. On souhaite construire un intervalle de confiance (exact) de p au niveau de confiance $1-\alpha$.

Etape 1 : appliquer l'inégalité de Bienaymé-Tchebychev.

Soit $\varepsilon > 0$. Soit $n \in \mathbb{N}^*$.

$$\mathbb{P}\left(\left[\left|\overline{X}_n - \mathbb{E}(\overline{X}_n)\right| > \varepsilon\right]\right) \leq \frac{\mathbb{V}(\overline{X}_n)}{\varepsilon^2}$$

et donc

$$\mathbb{P}\left(\left[\left|\overline{X}_n - p\right| > \varepsilon\right]\right) \leq \frac{p(1-p)}{n\varepsilon^2}$$

Etape 2 : fixer le niveau de risque.

On souhaite avoir $\dfrac{p(1-p)}{n\varepsilon^2} \leq \alpha$. Il faut choisir le paramètre ε pour que ce soit le cas. La première difficulté vient de la dépendance en p (on ne souhaite pas que ε dépende de p, car l'expression des estimateurs U_n et V_n ne doit pas dépendre de p).

L'étude de la fonction $x \mapsto x(1-x)$ sur $[0,1]$ donne : $p(1-p) \leq \dfrac{1}{4}$. On a donc

$$\frac{p(1-p)}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

et on résout ensuite l'équation

$$\begin{aligned}\frac{1}{4n\varepsilon^2} = \alpha &\iff 4n\varepsilon^2 = \frac{1}{\alpha}\\ &\iff \varepsilon^2 = \frac{1}{4n\alpha}\\ &\iff \varepsilon = \frac{1}{2\sqrt{n\alpha}}\end{aligned}$$

Etape 3 : expliciter l'intervalle de confiance obtenu.

D'après ce qui précède :

$$\mathbb{P}\left(\left[\left|\overline{X}_n - p\right| > \frac{1}{2\sqrt{n\alpha}}\right]\right) \leq \alpha$$

En passant au complémentaire, on obtient :

$$1 - \mathbb{P}\left(\left[\left|\overline{X}_n - p\right| \leq \frac{1}{2\sqrt{n\alpha}}\right]\right) \leq \alpha$$

i.e.

$$\mathbb{P}\left(\left[\left|\overline{X}_n - p\right| \leq \frac{1}{2\sqrt{n\alpha}}\right]\right) \geq 1 - \alpha$$

Or,

$$\left[\left|\overline{X}_n - p\right| \leq \frac{1}{2\sqrt{n\alpha}}\right] = \left[-\frac{1}{2\sqrt{n\alpha}} \leq p - \overline{X}_n \leq \frac{1}{2\sqrt{n\alpha}}\right] = \left[\overline{X}_n - \frac{1}{2\sqrt{n\alpha}} \leq p \leq \overline{X}_n + \frac{1}{2\sqrt{n\alpha}}\right]$$

Donc $\left[\overline{X}_n - \dfrac{1}{2\sqrt{n\alpha}}, \overline{X}_n + \dfrac{1}{2\sqrt{n\alpha}}\right]$ est un intervalle de confiance de p au niveau de confiance $1-\alpha$.

Remarque 8. Nous pouvons observer sur cet intervalle de confiance deux résultats intuitifs :

- plus le niveau de risque α souhaité est petit et plus l'intervalle de confiance est grand
- plus l'échantillon sondé est de taille n importante et plus l'intervalle de confiance est petit

Exemple 5. Précisons les observations précédentes en les quantifiant.

On note $\varepsilon = \dfrac{1}{2\sqrt{n\alpha}}$ la marge d'erreur (demi-amplitude de l'intervalle de confiance).

- Pour un échantillon de taille $n = 500$:

niveau de confiance $1-\alpha$ (en %)	70	75	80	85	90	95	97,5	99
marge d'erreur ε (en %)	4	4	5	6	7	10	14	22

- Pour un niveau de risque $\alpha = 5\%$:

taille de l'échantillon n	20	50	100	500	1000	2500	10000	50000
marge d'erreur ε (en %)	50	32	22	10	7	4	2	1

Le point de vue des instituts de sondage.

Lorsqu'un sondage est effectué, il faut systématiquement se poser la question des garanties aléatoires de précision sur lesquelles il se fonde. Pour les instituts de sondage, la question est donc de savoir combien de personnes il faut interroger pour obtenir un niveau de confiance $(1-\alpha)$ élevé et une précision importante (marge d'erreur ε faible).

Dans le tableau suivant, on calcule $n = \dfrac{1}{4\,\alpha\,\varepsilon^2}$ pour différentes valeurs du couple (ε, α) (exprimés en %).

ε \ $1-\alpha$	70	75	80	85	90	95	97,5	99
0,5	33333	40000	50000	66667	100000	200000	400000	1000000
1	8333	10000	12500	16667	25000	50000	100000	250000
1,5	3704	4444	5556	7407	11111	22222	44444	111111
2	2083	2500	3125	4167	6250	12500	25000	62500
2,5	1333	1600	2000	2667	4000	8000	16000	40000
3	926	1111	1389	1852	2778	5556	11111	27778
3,5	680	816	1020	1361	2041	4082	8163	20408
4	521	625	781	1042	1563	3125	6250	15625

- Le niveau de confiance $1-\alpha = 0,95$ est assez classique. Avec un tel niveau de confiance, on considère qu'il y a 95% de chances de tomber sur un panel standard. Lorsque c'est le cas, le paramètre cherché (ici p) se retrouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$.
  - On peut souhaiter obtenir un résultat très précis. Pour un sondage concernant des élections, savoir qu'un candidat est évalué à 19% plus ou moins $0,5\%$ serait idéal. Du point de vue du sondeur, cela voudrait dire interroger $n = 200000$ personnes. C'est inenvisageable pour des raisons évidentes de coût.
  - L'institut de sondage doit alors revoir ses objectifs à la baisse. En interrogeant $n = 3125$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 4%. Le coût est tout à fait envisageable mais le résultat semble alors un peu trop imprécis.
- Les résultats de ce dernier tableau démontrent que la méthode permettant d'obtenir un intervalle de confiance par inégalité de Bienaymé-Tchebychev est peu exploitable lorsque l'on cherche à obtenir des résultats relativement précis. Cela provient du fait que l'inégalité de Bienaymé-Tchebychev qui s'applique à toute v.a.r. (sans exploitation de la loi de celle-ci) est assez peu précise. Les instituts de sondage se basent sur une autre méthode consistant à obtenir un intervalle de confiance à l'aide du théorème central limite. Ce théorème énonce un résultat de convergence en loi. On sait que cette convergence se fait rapidement (des valeurs faibles de n fournissent de très bonnes approximations du résultat). En conséquence, on peut espérer obtenir des garanties aléatoires de précision fortes avec un nombre de sondés plus faible. C'est l'objet du paragraphe suivant.

3.2.2 Construction d'un intervalle de confiance asymptotique via le théorème central limite

Soit $\alpha \in]0,1[$. On souhaite construire un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$.

Etape 1 : appliquer le théorème central limite.

Les variables aléatoires $(X_i)_{i\in\mathbb{N}^*}$ sont indépendantes et suivent toutes la même loi. Elles admettent une espérance p et une variance $p(1-p)$ non nulle. On a donc, d'après le théorème central limite :

$$\overline{X_n}^* \xrightarrow[n\to+\infty]{\mathscr{L}} Z$$

où $\overline{X_n}^* = \sqrt{n}\dfrac{\overline{X}_n - p}{\sqrt{p(1-p)}}$ et $Z \hookrightarrow \mathcal{N}(0,1)$.

Ainsi, pour tout $(a,b) \in \mathbb{R}^2$ tel que $a \leq b$, on a

$$\mathbb{P}\left(\left[a \leq \sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \leq b\right]\right) \underset{n\to+\infty}{\longrightarrow} \Phi(b) - \Phi(a)$$

En particulier, pour tout $t \geq 0$,

$$\mathbb{P}\left(\left[\left|\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}\right| \leq t\right]\right) \underset{n\to+\infty}{\longrightarrow} \Phi(t) - \Phi(-t) = \Phi(t) - (1-\Phi(t)) = 2\Phi(t) - 1$$

Etape 2 : fixer le niveau de risque.

On résout l'équation

$$2\Phi(t) - 1 = 1 - \alpha \iff \Phi(t) = 1 - \frac{\alpha}{2} \iff t = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

On note alors $t_\alpha = \Phi^{-1}\left(1-\dfrac{\alpha}{2}\right)$ le *quantile* d'ordre $1-\dfrac{\alpha}{2}$.

> Pour $\alpha = 0,05$ (choix très courant, qui donne un intervalle de confiance de niveau de confiance 95%), on a : $1 - \dfrac{\alpha}{2} = 0,975$, donc $t_\alpha \approx 1,96$ (cf table de valeur de la loi normale centrée réduite).

Etape 3 : expliciter l'intervalle de confiance obtenu.

Tout d'abord, puisque $p(1-p) \leq \dfrac{1}{4}$:

$$\left[\left|\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}\right| \leq t_\alpha\right] = \left[\left|\overline{X}_n - p\right| \leq \frac{t_\alpha\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[\left|\overline{X}_n - p\right| \leq \frac{t_\alpha}{2\sqrt{n}}\right]$$

D'où

$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\overline{X}_n - \frac{t_\alpha}{2\sqrt{n}} \leq p \leq \overline{X}_n + \frac{t_\alpha}{2\sqrt{n}}\right]\right) \geq \lim_{n\to+\infty} \mathbb{P}\left(\left[\left|\sqrt{n}\frac{\overline{X}_n - p}{\sqrt{p(1-p)}}\right| \leq t_\alpha\right]\right) = 1-\alpha$$

Donc $\left[\overline{X}_n - \dfrac{t_\alpha}{2\sqrt{n}}, \overline{X}_n + \dfrac{t_\alpha}{2\sqrt{n}}\right]$ est un intervalle de confiance asymptotique de p au niveau de confiance $1-\alpha$.

Exemple 6. On note $\varepsilon = \dfrac{t_\alpha}{2\sqrt{n}}$ la marge d'erreur (demi-amplitude de l'intervalle de confiance asymptotique).

- Pour un échantillon de taille $n = 500$:

niveau de confiance $1-\alpha$ (en %)	70	75	80	85	90	95	97,5	99
marge d'erreur ε (en %)	2,3	2,6	2,9	3,2	3,7	4,4	5,1	5,7

- Pour un niveau de risque $\alpha = 5\%$:

taille de l'échantillon n	20	50	100	500	1000	2500	10000	50000
marge d'erreur ε (en %)	21.9	13.9	9.8	4.4	3.1	1.96	1.0	0.4

Le point de vue des instituts de sondage.

Lorsqu'un sondage est effectué, il faut systématiquement se poser la question des garanties aléatoires de précision sur lesquelles il se fonde. Pour les instituts de sondage, la question est donc de savoir combien de personnes il faut interroger pour obtenir un niveau de confiance $(1-\alpha)$ élevé et une précision importante (marge d'erreur ε faible).

Dans le tableau suivant, on calcule $n = \dfrac{t_\alpha^2}{4\,\varepsilon^2}$ pour différentes valeurs du couple (ε, α) (exprimés en %).

ε \ $1-\alpha$	70 (1,04)	75 (1,17)	80 (1,28)	85 (1,44)	90 (1,64)	95 (1,96)	97,5 (2,26)	99 (2,57)
0,5	10816	13689	16384	20736	26896	38416	51076	66049
1	2704	3422	4096	5184	6724	9604	12769	16512
1,5	1202	1521	1820	2304	2988	4268	5674	7339
2	676	856	1024	1296	1681	2401	3192	4128
2,5	433	548	655	829	1076	1537	2043	2642
3	300	380	455	576	747	1067	1419	1835
3,5	221	279	334	423	549	784	1042	1348
4	169	214	256	324	420	600	798	1032

- Sur la première ligne, on a placé entre parenthèse la valeur de t_α correspondante au niveau de confiance $1-\alpha$ considéré. Par exemple, si $1-\alpha = 0,95$ alors $1-\dfrac{\alpha}{2} = 0,975$ et $t_\alpha \simeq 1,96$.
- Comme mentionné précédemment, le niveau de confiance $1-\alpha = 0,95$ est assez classique. Avec un tel niveau de confiance, on considère qu'il y a 95% de chances de tomber sur un panel standard. Lorsque c'est le cas, le paramètre réel se retrouve dans l'intervalle $[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon]$.
  - On peut souhaiter obtenir un résultat très précis. Pour un sondage concernant des élections, savoir qu'un candidat est évalué à 19% plus ou moins 0,5% serait idéal. Du point de vue du sondeur, cela voudrait dire interroger $n = 38416$ personnes. C'est 5 fois moins que pour l'intervalle de confiance obtenu par inégalité de Bienaymé-Tchebychev. Pour autant, c'est toujours inenvisageable pour des raisons de coût.
  - L'institut de sondage doit alors revoir ses objectifs à la baisse. En interrogeant $n = 2401$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 2%. C'est 2 fois moins de marge d'erreur que dans le cas de l'intervalle de confiance obtenu par inégalité de Bienaymé-Tchebychev. Le coût est tout à fait envisageable et le résultat offre une précision correcte.
- Notons que les intervalles de confiance obtenus par les deux méthodes ont été réalisés avec la majoration : $p(1-p) \leq \dfrac{1}{4}$ $(*)$.
  - La valeur $\frac{1}{4}$ est atteinte dans le cas où $p = \frac{1}{2}$. La majoration $(*)$ est donc la meilleure que l'on puisse faire en l'absence d'information sur p.
  - Le rôle d'un sondage est justement d'obtenir de l'information sur p (une valeur approchée). Si un candidat est évalué à 20% (resp. 80%) alors on a $p(1-p) \simeq 0,16$.
    Avec ce calcul et pour $1-\alpha = 0,95$ et $n = 1500$, on obtient alors :

$$\varepsilon = \frac{\sqrt{p(1-p)}}{\sqrt{n}}\, t_\alpha \simeq \frac{\sqrt{0,16}}{\sqrt{1500}}\, 1,96 \simeq 0,02$$

    Ainsi, les sondages font souvent valoir une marge d'erreur qui dépend de l'estimation du candidat.

3.2.3 Simulations informatiques

Exercice 8 : Compléter la fonction **Python** qui suit pour qu'elle simule un sondage fait sur un échantillon de n personnes afin d'obtenir un intervalle de confiance de p au niveau de confiance $1-\alpha$ en utilisant l'inégalité de Bienaymé-Tchebychev. Le paramètre p sera choisi aléatoirement au début de la fonction.

```
def simulationSondageBT(n, alpha):
    p = _____________ # Paramètre à estimer
    sondage = _____________________ # Résultats du sondage
    Xbar = ____________________
    eps = ____________________________
    u = Xbar - eps
    v = Xbar + eps
    if _____________:
        print('Intervalle de confiance valide')
    else:
        print('Intervalle de confiance non valide')
    return p, [u,v]
```

Exercice 9 : Compléter la fonction **Python** qui suit pour qu'elle simule un sondage fait sur un échantillon de n personnes afin d'obtenir un intervalle de confiance de p au niveau de confiance 95% en utilisant le théorème central limite. Le paramètre p sera choisi aléatoirement au début de la fonction.

```
def simulationSondageTCL(n, alpha):
    p = _____________ # Paramètre à estimer
    sondage = _____________________ # Résultats du sondage
    Xbar = ____________________
    t = _________
    eps = ____________________________
    u = Xbar - eps
    v = Xbar + eps
    if _____________:
        print('Intervalle de confiance valide')
    else:
        print('Intervalle de confiance non valide')
    return p, [u,v]
```

3.3 Intervalle de confiance asymptotique avec variance inconnue

On introduit la *variance empirique* obtenue à partir de la moyenne empirique

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

On peut utiliser l'estimateur ci-dessus pour former un intervalle de confiance pour l'espérance lorsque la variance est aussi inconnue.

Théoreme 1. *Soit X une variable aléatoire d'espérance m et de variance non nulle inconnue et $(X_1, \dots, X_n)$ un n-échantillon de X. Alors, l'intervalle*

$$\left[\overline{X}_n - t_\alpha \frac{\overline{S}_n}{\sqrt{n}}, \overline{X}_n + t_\alpha \frac{\overline{S}_n}{\sqrt{n}}\right],$$

où $\Phi(t_\alpha) = 1 - \alpha/2$, est un intervalle de confiance asymptotique pour m au risque α.

4 Exercices supplémentaires

4.1 Un exemple historique. Le *German Tank Problem*

À partir des numéros de série observés sur des tanks ennemis, peut-on *estimer* le nombre total d'engins des forces adverses ?

Le contexte. Été 1943, les Alliés essaient de percer le bloc de l'Axe en créant un nouveau front via l'Italie. Ils rencontrent un nouveau type de char allemand, le bien nommé *Sonderkraftfahrzeug 171* plus connu des afficionados de machines de combat sous le nom de *Panther*.

Ce dernier est mieux équipé et plus performant que ceux rencontrés jusqu'alors. Il a été conçu en réponse à l'excellent *T-34* utilisé par les soviétiques sur le front de l'Est. Il peut percer les défenses et détruire la majorité des tank alliés.

Néanmoins, malgré sa puissance théorique, celui-ci ne peut avoir un réel impact sur l'issue de la guerre que si le nombre d'unités produites est suffisant. Il apparait crucial pour les Alliés de déterminer ou plutôt d'*estimer* combien de *Panther* étaient produits. La tâche fut confiée à [la] *Economic Warfare Division of the American Embassy in London*[1].

La modélisation. On suppose que l'ennemi produit une série de chars immatriculés par des entiers en commençant par 1. En plus de cela, quelle que soit la date de production du char, ses années de service, ou encore son numéro de série, la distribution des numéros d'immatriculation est considérée comme étant uniforme dès l'instant où on mène l'analyse.

Dans notre modélisation, les allemands disposent de N tanks numérotés de 1 à N. Les force alliées observent aléatoirement, uniformément et "avec remise" n numéros de séries $(X_1, \dots, X_n)$ et cherchent à estimer le paramètre N.

On considère dans tout le problème un n-échantillon $(X_1, \dots, X_n)$ de la loi $\mathcal{U}(\llbracket 1, N \rrbracket)$ et une première idée serait de considérer la *moyenne empirique* des valeurs observées, on commence donc par poser

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

1. Que vaut $\mathbb{E}(\overline{X}_n)$? Il serait *pratique* qu'en moyenne, la variable aléatoire choisie pour estimer N renvoie N. Expliciter alors un estimateur T_n de N, fonction du n-échantillon $(X_1, \dots, X_n)$, tel que $\mathbb{E}(T_n) = N$. On dit dans ce cas que T_n est *sans biais*.
2. Calculer $\mathbb{V}(T_n)$ et montrer, à l'aide de l'inégalité de Bienaymé-Tchebychev, que
$$\forall \varepsilon > 0, \qquad \lim_{n \to +\infty} \mathbb{P}\left([|T_n - N| > \varepsilon]\right) = 0.$$
3. Ce résultat semble affirmer que l'estimateur T_n converge (dans un certain sens) vers N, c'est à dire que si n est assez grand (si on dispose de suffisamment de données) l'estimation de N obtenue comme réalisation de T_n est proche de N. En revanche, imaginons qu'on ait 5 données $D = [8, 322, 15, 135, 69]$, que vaut l'estimation obtenue avec T_5 correspondant à cette observation ? Ce résultat incite à introduire un autre estimateur, qui serait plus performant pour de plus petites valeurs de n.

On introduit alors le nouvel estimateur $M_n = \max(X_1, \dots, X_n)$.

4. Calculer, pour tout $k \in \mathbb{N}$, $\mathbb{P}([M_n \leq k])$.
5. Soit Y une variable aléatoire à valeurs dans $\llbracket 1, N \rrbracket$. Montrer que
$$\mathbb{E}(Y) = \sum_{k=0}^{N-1} \mathbb{P}([Y > k]).$$

1. Comme le raconte l'article *An Empirical Approach to Economic Intelligence in World War II*, R. Ruggles & H. Brodie, Journal of the American Statistical Association **42**-237 (1947), 72-91

6. Montrer alors que :
$$\mathbb{E}(M_n) = N - \sum_{k=0}^{N-1} \left(\frac{k}{N}\right)^n.$$

7. Vérifier que, pour tout $k \in [\![0, N-1]\!]$:
$$0 \leq \left(\frac{k}{N}\right)^n \leq N \int_{k/N}^{(k+1)/N} t^n \mathrm{d}t.$$

8. En déduire que :
$$N - \frac{N}{n+1} \leq \mathbb{E}(M_n) \leq N$$
puis que :
$$\lim_{n\to+\infty} \mathbb{E}(M_n) = N.$$
(On dit dans ce cas que l'estimateur M_n est *asymptotiquement sans biais*.)

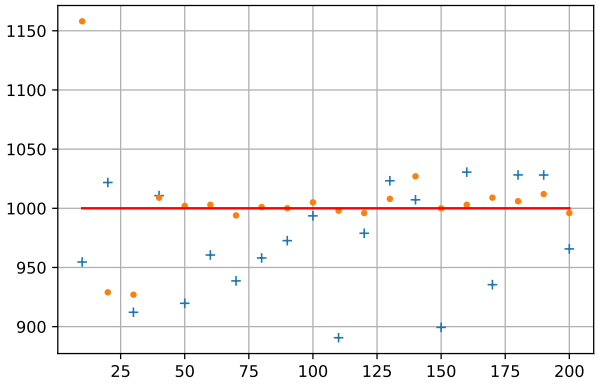
Si l'estimateur M_n parait naturel, il a clairement un défaut ; il sous-estime nécessairement N (puisqu'il renverra toujours une valeur inférieure (ou égale) à N). On va donc essayer d'y apporter une légère *correction*.

Commençons par introduire le numéro du plus petit tank observé $m_n = \min(X_1, \ldots, X_n)$.

Comme N est inconnu, on ne connait pas l'écart entre N et M_n, mais il parait raisonnable de penser qu'il y a (en moyenne) autant de tanks *non observés* entre M_n et N qu'entre 1 et m_n. Entre le plus petit numéro observé et le tank avec le numéro de série 1, il y a $m_n - 1$ numéros de tank. On pense alors à ajouter la correction
$$\tilde{M}_n = M_n + (m_n - 1).$$

9. En s'inspirant des calculs précédents pour M_n, déterminer $\mathbb{E}(m_n)$ sous forme d'une somme qu'on ne cherchera pas à simplifier.
10. Montrer que $\tilde{M}_n$ vérifie maintenant $\mathbb{E}(\tilde{M}_n) = N$.
11. **Comparaison des estimateurs**. On propose le script **Python** ci-dessous. Quel estimateur semble le plus performant ?

```
def estimateurs(N, n):
    X= rd.randint(1, N+1, n)
    T = 2*np.mean(X)-1
    M = np.max(X)+np.min(X)-1
    return [T,M]
T = [ ]; M = [ ]; N = 1000
Xabscisse = [10*i for i in range(1, 21)]
for n in Xabscisse:
    [t,m] = estimateurs(N,n)
    T.append(t)
    M.append(m)
plt.plot(Xabscisse, T, '+')
plt.plot(Xabscisse, M, '.')
plt.plot(Xabscisse, [N for k in x], 'red')
plt.show()
```



12. Choisir alors l'estimateur le plus performant pour proposer une estimation du nombre de tanks ennemis à partir des données top secrètes transmises par les services de renseignement, au péril de leur vie.

```
X=[14, 44, 50, 101, 117, 127, 134, 139, 165, 188, 192, 201, 204, 215,
   234, 243, 244, 253, 269, 269, 282, 287, 288, 322, 345]
```

4.2 Maximum de vraisemblance

Exercice 10 : Un jeu télévisé consiste à poser à un candidat une succession de questions à choix multiples. Les questions sont posées dans un ordre de difficulté croissant et rapportent de plus en plus d'argent au candidat. L'équipe qui conçoit les questions décide de tester la difficulté d'une d'entre elles pour savoir à quel moment du jeu il serait préférable de la poser. Pour ce faire, on se propose de réaliser un sondage dans la population.

Modélisation du problème

- On propose à chaque personne interrogée 3 réponses, la réponse correcte étant la réponse 1.
- Le comportement d'une personne interrogée est le suivant :
  - si elle connaît la réponse correcte, elle la donne.
  - sinon elle choisit au hasard une des **trois** réponses proposées. On prend ainsi en compte la possibilité qu'une personne interrogée donne la réponse correcte par chance.
- On note X la v.a.r. égale à la réponse donnée par la personne interrogée. On note Y la v.a.r. égale à 1 si la personne interrogée connaît la bonne réponse et à 0 sinon. Enfin, on note θ (paramètre que l'on cherche à estimer) la probabilité qu'une personne de la population **connaisse** la réponse correcte.

1. (a) Reconnaître la loi de Y.
   (b) Déterminer, en fonction de θ, la loi de X. On note $p = \mathbb{P}([X = 1])$. Exprimer alors θ en fonction de p.
   (c) Quelle est, en fonction de θ, la probabilité qu'une personne ayant choisi la réponse 1 l'ait fait car elle connaissait réellement la réponse ?
2. Afin d'estimer θ, on constitue dans la population n groupes de 30 personnes qui seront interrogées par un enquêteur. Pour $1 \leq i \leq n$, on note V_i la variable égale au nombre de réponses 1 obtenues dans le groupe i. Les v.a.r. V_i sont supposées mutuellement indépendantes. On note enfin $Z_n = \dfrac{V_1 + \cdots + V_n}{30n}$.
   (a) Déterminer l'espérance de Z_n, et sa variance.
   (b) Déterminer, à partir de Z_n, un estimateur sans biais T_n de θ.
   (c) Déterminer le risque quadratique de T_n.
   (d) Montrer : $\forall \varepsilon > 0,\ \mathbb{P}([|T_n - \theta| \geq \varepsilon]) \leq \dfrac{1}{20\ n\ \varepsilon^2}$. Que peut-on en déduire sur l'estimateur T_n ?
3. Dans la question précédente, on a proposé un estimateur T_n de θ. L'estimateur initial Z_n, biaisé, n'a pas été retenu mais a permis de construire l'estimateur sans biais T_n. Une estimation $\hat{\theta}$ de θ est alors fournie par une réalisation de T_n. Dans cette question, on cherche à obtenir une estimation $\hat{p}$ de p. Pour ce faire, on va raisonner comme suit : on part d'une réalisation $(v_1, v_2, \ldots, v_n)$ de l'échantillon $(V_1, V_2, \ldots, V_n)$ et on cherche à obtenir, grâce à cette donnée, le meilleur estimateur pour p. On s'intéresse alors à la quantité $L(p)$ suivante :

$$L(p) \ = \ \mathbb{P}_p([V_1 = v_1] \ \cap \ \ldots \ \cap \ [V_n = v_n])$$

   L est une fonction appelée **vraisemblance**. Elle permet de mesurer la probabilité que notre modèle ait donné lieu à l'observation $(v_1, \ldots, v_n)$. Le principe du **maximum de vraisemblance** est de choisir comme estimation de p la valeur qui maximise la vraisemblance de modèle par rapport à la donnée $(v_1, \ldots, v_n)$.
   (a) Expliciter, en fonction de p, la valeur de $L(p)$.
   (b) Étudier les variations de la fonction $f : p \mapsto \ln(L(p))$.
   (c) Montrer que f et donc L admet un maximum. On note $\hat{p}$ le point en lequel f atteint ce maximum.
   (d) Déterminer l'estimateur du maximum de vraisemblance.

4.3 Intervalles de confiance

Exercice 11 : Soit $(X_1, \dots, X_n)$ un n-échantillon d'une variable aléatoire $X \hookrightarrow \mathcal{U}([0, \theta])$ où $\theta > 0$. On a l'estimateur sans biais de θ :

$$V_n = \frac{2}{n}(X_1 + \cdots + X_n)$$

1. Montrer, à l'aide de l'inégalité de Bienaymé-Tchébychev, que pour tout $\varepsilon > 0$:

$$\mathbb{P}([|V_n - \theta| > \varepsilon]) \leq \frac{\theta^2}{3n\varepsilon^2}.$$

2. Soit $\alpha \in]0, 1[$. Montrer que, pour n assez grand,

$$\left[\theta \in \left[V_n - \sqrt{\frac{\theta^2}{3n\alpha}}, V_n + \sqrt{\frac{\theta^2}{3n\alpha}}\right]\right] = \left[\frac{V_n}{1 + \frac{1}{\sqrt{3n\alpha}}} \leq \theta \leq \frac{V_n}{1 - \frac{1}{\sqrt{3n\alpha}}}\right]$$

3. En déduire un intervalle de confiance au risque α pour θ.

Exercice 12 : Le second tour d'une élection met en présence deux candidats A et B. On souhaite réaliser un sondage afin de connaître, avec un niveau de confiance de $0,95$, le futur vainqueur. Sachant par ailleurs que les deux candidats sont au coude à coude, on veut réduire la marge d'erreur à $0,01$.

1. Donner le nombre minimal d'électeurs à interroger si on se fie à l'inégalité de Bienaymé-Tchebychev pour faire le calcul.
2. Même question en utilisant le théorème central limite.

Exercice 13 : Une entreprise souhaite acquérir une machine qui fabrique un certain type d'objets et qui, en fonctionnement normal, produit une proportion p $(0 < p < 1)$ d'objets défectueux. Le directeur veut connaître la valeur de p. Pour cela, il teste la machine et prélève un échantillon de n objets qu'il analyse, avec $n \geq 1$. Pour tout $i \in [\![1, n]\!]$, soit X_i la v.a.r. de Bernoulli définie par :

$$X_i = \begin{cases} 1 & \text{si le } i^{ième} \text{ objet prélevé est défectueux} \\ 0 & \text{sinon} \end{cases}$$

On suppose que dans les conditions de prélèvement, les variables aléatoires $X_1, \dots, X_n$ sont indépendantes.

On pose $S_n = \sum_{k=1}^{n} X_k$.

1. (a) Montrer que $F_n = \frac{S_n}{n}$ est un estimateur sans biais de p.
   (b) Calculer le risque quadratique r_n de F_n. Déterminer $\lim\limits_{n \to +\infty} r_n$.
2. Soit α un réel de $]0, 1[$. On souhaite déterminer dans cette question un intervalle de confiance du paramètre p inconnu, au niveau de confiance $1 - \alpha$, à partir de l'échantillon $(X_1, \dots, X_n)$.
   (a) Quelle est la limite en loi de la suite $\left(\sqrt{n}\frac{F_n - p}{\sqrt{p(1-p)}}\right)_{n \in \mathbb{N}^*}$?
   (b) Soit f_n la réalisation de F_n sur l'échantillon considéré. Soit t_α le réel défini par $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$, où Φ désigne la fonction de répartition de la loi normale centrée, réduite. Montrer qu'un intervalle de confiance de p au niveau $1 - \alpha$ est donné par $[U_n, V_n]$ où :

$$U_n = F_n - \frac{t_\alpha}{2\sqrt{n}} \quad \text{et} \quad V_n = F_n + \frac{t_\alpha}{2\sqrt{n}}$$

   (c) On suppose dans cette question qu'en fonctionnement normal la machine produit une proportion $p = 0,05$ d'objets défectueux. Le directeur analyse 10 000 objets et compte 600 objets défecteux sur cet échantillon. Décide-t-il d'acheter la machine, au niveau de confiance de 95% ? On donne $\Phi(2) \simeq 0,975$.

Exercice 14 : Soit X une v.a.r. suivant la loi uniforme sur $[0,\theta]$, où $\theta > 0$ est un paramètre inconnu. Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon de la v.a.r. X. On considère les variables aléatoires suivantes :

$$U_n = \max(X_1, X_2, \dots, X_n) \quad \text{et} \quad T_n = n\left(1 - \frac{U_n}{\theta}\right)$$

On souhaite déterminer un intervalle de confiance asymptotique du paramètre θ de la forme $[U_n, V_n]$, au niveau de confiance $1-\alpha$.

1. T_n peut-il être un estimateur de θ ?
2. Déterminer la fonction de répartition F_{U_n} de la variable U_n. En déduire la fonction de répartition F_{T_n} de la variable T_n.
3. Prouver que $(T_n)_{n\in\mathbb{N}^*}$ converge en loi vers une v.a.r. T suivant la loi $\mathcal{E}(1)$.
4. Montrer l'égalité des événements $[U_n \leq \theta \leq V_n]$ et $\left[0 \leq T_n \leq n\left(1 - \frac{U_n}{V_n}\right)\right]$.
5. En déduire que l'intervalle cherché est obtenu pour :
$$V_n = \frac{U_n}{1 + \frac{1}{n}\ln(\alpha)}$$
6. On considère le programme suivant :

```
1 n = int(input('Valeur de n ?'))
2 theta = 5 * rd.random()
3 for i in range(n):
4     print(rd.uniform(0,theta))
```

Une exécution de ce programme affiche les nombres suivants :

0.8608569	0.1431483	0.9570818	0.8822904	0.1341774
1.0237293	0.9650951	0.2335499	0.6681662	0.3256168

(a) On considère un niveau de confiance de $0,95$ ($\ln(0,05) \simeq -3$). Déduire des valeurs précédentes les réalisations de u_n et v_n correspondantes. Quel est l'intervalle de confiance observé correspondant ?

(b) Quelle valeur faut-il donner à n pour avoir $V_n = 1,01 \times U_n$?

Exercice 15 : (HEC 2008) Dans tout le problème, N désigne un entier naturel fixé supérieur ou égal à 2, et p un réel fixé de l'intervalle $]0,1[$. On pose $q = 1-p$. Soit n un entier naturel quelconque. Dans une population de N individus, on s'intéresse à la propagation d'un certain virus. Chaque jour, on distingue dans cette population trois catégories d'individus : en premier lieu, les individus sains, c'est-à-dire ceux qui ne sont pas porteur du virus, ensuite les individus qui viennent d'être contaminés et qui sont innofensifs pour les autres, et enfin, les individus contaminés par le virus et qui sont contagieux. Ces trois catégories évoluent jour après jour selon le modèle suivant :

1. chaque jour n, chaque individu sain peut être contaminé par n'importe lequel des individus contagieux avec la même probabilité p, ces contaminations éventuelles étant indépendantes les unes des autres ;
2. un individu contaminé le jour n devient contagieux le jour $n+1$;
3. chaque individu contagieux le jour n redevient sain le jour $n+1$.

On suppose que le paramètre p, qui exprime la probabilité qu'un individu contagieux transmette le virus à un individu sain, est inconnu, et on cherche à l'estimer. On rappelle que : $q = 1-p$. Pour m entier supérieur ou égal à 1, on considère un m-échantillon $(Y_1, Y_2, \dots, Y_m)$ de variables aléatoires indépendantes, de même loi de Bernoulli de paramètre p, définies sur un même espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$. On pose : $\overline{Y_m} = \dfrac{1}{m}\sum_{k=1}^{m} Y_k$. *Dans toute cette partie, on note ε un réel strictement positif quelconque.*

1. (a) Montrer que $\overline{Y_m}$ est un estimateur de p.

   (b) A l'aide de l'inégalité de Bienaymé-Tchebycheff, montrer que l'intervalle $\left[\overline{Y_m} - \sqrt{\dfrac{5}{m}}, \overline{Y_m} + \sqrt{\dfrac{5}{m}}\right]$ est un intervalle de confiance de p au niveau de confiance 0.95.

2. Soit θ un réel strictement positif.

   (a) Etablir l'égalité suivante :
   $$\mathbb{P}\left(\left[\overline{Y_m} - p \geq \varepsilon\right]\right) = \mathbb{P}\left(\left[\mathrm{e}^{m\theta\overline{Y_m}} \geq \mathrm{e}^{m\theta(p+\varepsilon)}\right]\right)$$

   (b) Montrer que si T est une v.a.r. discrète finie à valeurs positives d'espérance $\mathbb{E}(T)$, et a un réel strictement positif, on a l'inégalité :
   $$\mathbb{P}([T \geq a]) \leq \frac{\mathbb{E}(T)}{a}$$

   (c) Soit g la fonction définie sur $\mathbb{R}_+$ par : $g(x) = \ln(p\,\mathrm{e}^x + q)$. Déduire des questions précédentes l'inégalité suivante :
   $$\mathbb{P}([\overline{Y_m} - p \geq \varepsilon]) \leq \mathrm{e}^{m(g(\theta)-\theta(p+\varepsilon))}$$

   (d) Montrer que la fonction g est de classe $\mathcal{C}^2$ sur $\mathbb{R}_+$ et vérifie, pour tout x de $\mathbb{R}_+$, l'inégalité : $\left|g''(x)\right| \leq \dfrac{1}{4}$.

   (e) En déduire l'inégalité suivante : $g(\theta) \leq \theta p + \dfrac{\theta^2}{8}$.

   (f) Étudier les variations de la fonction $h : x \mapsto \dfrac{x^2}{8} - \varepsilon x$ sur $\mathbb{R}_+$. En déduire l'inégalité : $\mathbb{P}([\overline{Y_m} - p \geq \varepsilon]) \leq \mathrm{e}^{-2m\varepsilon^2}$.

3. On pose $\overline{W_m} = \dfrac{1}{m}\sum_{k=1}^{m}(1-Y_k)$. Établir l'inégalité : $\mathbb{P}([\overline{W_m} - q \geq \varepsilon]) \leq \mathrm{e}^{-2m\varepsilon^2}$.

4. (a) Déduire des questions 2.f) et 3, l'inégalité suivante :
   $$\mathbb{P}([\left|\overline{Y_m} - p\right| \geq \varepsilon]) \leq 2\mathrm{e}^{-2m\varepsilon^2}$$

   (b) Sachant $\ln(0.025) \simeq -3.688$, calculer $2\mathrm{e}^{-2m\varepsilon^2}$ pour $\varepsilon = \sqrt{\dfrac{1.844}{m}}$. En déduire un nouvel intervalle de confiance de p au niveau de confiance 0.95. Comparer cet intervalle de confiance avec celui obtenu à la question 1.b). Conclure.

Exercice 16 : Le but de ce problème est l'étude d'estimateurs du nombre N d'individus d'une population. Une réserve naturelle contient N oiseaux. Le nombre N est inconnu. On capture au hasard m oiseaux dans la réserve, on les bague et on les relâche. Posons $p = \dfrac{m}{N}$ la proportion des oiseaux de la population qui sont bagués. On a : $0 < m < N$, où m et N sont deux entiers.

Partie I. On choisit successivement au hasard, avec remise, n oiseaux dans la population. On appelle I_n le nombre d'oiseaux bagués obtenus lors de ces n choix.

1. Quelle est la loi de I_n ? Donner son espérance et sa variance en fonction de n et de p.
2. Justifier que $\dfrac{1}{nm} I_n$ est un estimateur sans biais de $\dfrac{1}{N}$.
3. Montrer que $\dfrac{1}{nm} I_n$ est convergent, c'est-à-dire que pour tout $\varepsilon > 0$,
$$\lim_{n\to+\infty} \mathbb{P}\left(\left[\left|\frac{1}{nm} I_n - \frac{1}{N}\right| > \varepsilon\right]\right) = 0$$
4. Dans cette question, on suppose que $n = 1\,600$ et que l'on a eu 400 oiseaux bagués pami les 1 600 choisis.
   (a) Déterminer, à l'aide de l'estimateur I_n, un estimateur sans biais de p.
   (b) Déterminer un intervalle de confiance de p au taux de confiance de $0,95$. On donne $\Phi(2) = 0,975$.
   (c) Sachant que l'on a marqué 990 oiseaux, en déduire un intervalle de confiance de N avec un risque d'erreur d'au plus 5%.
5. On pose $Y_n = \dfrac{m(n+1)}{I_n + 1}$. *(on ne peut pas prendre* $\dfrac{nm}{I_n}$ *car* I_n *peut prendre la valeur* 0*)*
   (a) Montrer que $\mathbb{E}(Y_n) = N\left(1 - (1-p)^{n+1}\right)$. On pourra utiliser l'égalité : $\dfrac{1}{k+1}\binom{n}{k} = \dfrac{1}{n+1}\binom{n+1}{k+1}$.
   (b) Y_n est-il un estimateur sans biais de N ?
   (c) Montrer que l'estimateur Y_n est asymptotiquement sans biais, c'est-à-dire que $\lim\limits_{n\to+\infty} \mathbb{E}(Y_n) = N$.

Partie II. On choisit au hasard et avec remise des oiseaux de la population. On appelle R_n le nombre de choix effectués pour obtenir n oiseaux bagués. Ainsi, R_n est la rang de sortie du $n^{ième}$ oiseau bagué dans la suite des choix.

6. Quelle est la loi de R_1 ? Donner l'expression de $\mathbb{P}([R_1 = k])$, en fonction de k et de p. Donner l'espérance de R_1 et sa variance.
7. On pose $D_1 = R_1$ et pour tout entier k tel que $k \geq 2$, $D_k = R_k - R_{k-1}$. Ainsi, D_k est le nombre de choix effectués après l'obtention du $(k-1)^{ième}$ oiseau bagué pour obtenir le $k^{ième}$.
   (a) Justifier que les variables $D_1, D_2, \ldots, D_k, \ldots$ sont mutuellement indépendantes, et suivent la même loi que R_1.
   (b) Soit n un entier supérieur ou égal à 2. Calculer R_n en fonction de D_k, pour $1 \leq k \leq n$. En déduire l'espérance et la variance de R_n en fonction de n et de p.
8. On pose $X_n = \dfrac{m}{n} R_n$. Montrer que X_n est un estimateur sans biais.
9. (a) À l'aide de quel théorème peut-on affirmer que l'on peut approcher la loi de $\dfrac{p R_n - n}{\sqrt{n(1-p)}}$ par la loi normale centrée réduite, pour n suffisamment grand ? Montrer qu'alors la loi de X_n peut, elle aussi, être approchée par une loi normale dont on donnera les valeurs des paramètres.
   (b) On suppose dans cette question que X_n suit une loi normale, que $m = 1000$ et que $p \geq 0,2$. Déterminer une valeur de n à partir de laquelle on peut affirmer que l'on connaît N à 500 près avec une probabilité d'au moins $0,95$. On utilisera l'approximation $\Phi(2) \simeq 0,975$.
10. Posons C_k l'événement : « le $k^{ième}$ choix est celui d'un oiseau bagué ». Exprimer $[R_n = k]$ à l'aide de la variable I_{k-1} définie dans la partie I et de l'événement C_k. Puis en déduire la loi de R_n.
11. En déduire que si $x \in]0,1[$, alors la série $\sum\limits_{i\geq 0} \binom{n+i}{n} x^i$ est convergente, et calculer sa somme.