# Table des matières

# 1 Introduction : un exemple historique

Dans tous les chapitres précédents concernant les probabilités, les lois des processus aléatoires étaient connues, et c'est à partir de ces lois que nous avons fait les calculs.

Adoptons alors une approche différente d'ordre pratique ; on **ne connait pas** la loi de $X$. Par exemple la durée de vie d'un composant électronique, le nombre de clients se présentant au guichet d'une banque un jour donné, la proportion d'un certain caractère dans une population, le résultat d'un scrutin à partir des premiers dépouillements...

Il peut être raisonnable de penser qu'elle appartient à une famille de lois usuelles (ce qui définit le **modèle** de loi) qui dépend d'un paramètre (inconnu) qu'on aimerait donc **estimer** à partir du résultat $(x_1, x_2, \dots, x_n)$ de $n$ expériences.

Il s'agit bien d'une estimation car on ne peut pas faire une infinité d'expérience de même qu'on ne peut pas interroger l'ensemble de la population.

Ainsi, le $n$-uplet $(x_1, x_2, \dots, x_n)$ correspond alors à l'**observation** d'un **échantillon**, c'est à dire à sa réalisation pour une certaine issue $\omega \in \Omega$. Il s'agit de données dont on dispose. On peut alors, à partir des observations, vouloir produire une valeur approchée du paramètre (par exemple de l'espérance), ce qui s'appelle une *estimation ponctuelle* (à l'aide d'une *formule* bien choisie appliquée aux données) ou vouloir fournir un intervalle qui contient le paramètre avec un haut niveau de confiance, ce qu'on appelle *estimation par intervalle de confiance*.

Ce chapitre clôt le cours de *Mathématiques appliquées* de cette année. Ainsi, il est de bon goût de le commencer par une application *concrète* de la théorie des statistiques, qui devrait motiver l'introduction des notions qui suivront.

## 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. 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

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 une variable aléatoire $T_n$, fonction du $n$-échantillon $(X_1, \dots, X_n)$ telle 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) la valeur approchée de $N$ obtenue avec la définition de $T_n$ appliquée à l'observation 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 ? Que cela motive-t-il ?

On introduit alors le nouvel *estimateur*, c'est à dire une nouvelle fonction du $n$-échantillon
$$M_n = \max(X_1, \dots, X_n)$$

4. Calculer, pour tout $k \in \mathbb{N}$, $\mathbb{P}([M_n \leq k])$.
5. Soit $Y$ une variable aléatoire à valeurs dans $[\![1, N]\!]$. Montrer que
$$\mathbb{E}(Y) = \sum_{k=0}^{N-1} \mathbb{P}([Y > k]).$$
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 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 valeurs 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 $m_n = \min(X_1, \dots, X_n)$.

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
$$\tilde{M}_n = M_n + (m_n - 1).$$

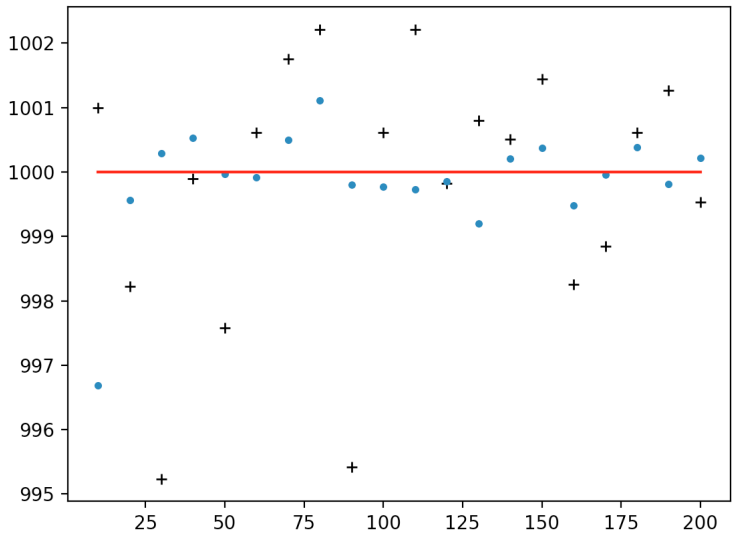
9. En s'inspirant des calculs précédents pour $M_n$, déterminer $\mathbb{E}(m_n)$ sous forme d'une somme qu'on ne cherchera pas à simplifier.
10. Montrer que $\tilde{M}_n$ vérifie maintenant $\mathbb{E}(\tilde{M}_n) = N$.

11. **Comparaison des estimateurs**. On propose la fonction Python ci-dessous. Que fait-elle ?

```
import numpy as np
import numpy.random as rd
def mystere(N, n):
    T=[]
    M=[]
    for j in range(1000):
        X=[rd.randint(1, N+1) for k in range(n)]
        T.append(2*np.mean(X)-1)
        M.append(np.max(X)+np.min(X)-1)
    t=np.mean(T)
    m=np.mean(M)
    return [t,m]
```

On ajoute les instructions ci-dessous dont l'exécution permet d'obtenir la figure ci-contre. Interpréter. Quel estimateur semble le plus performant ?

```
import matplotlib.pyplot as plt
T=[ ]
M=[ ]
N=1000
x=[10*i for i in range(1, 21)]
for n in x:
    [t,m]=mystere(N,n)
    T.append(t)
    M.append(m)
plt.plot(x, T, '+')
plt.plot(x, M, '.')
plt.plot(x, [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 service 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]
```

# 2 Estimation ponctuelle

On considère un phénomène aléatoire et on s'intéresse à une variable aléatoire $X$ qui lui est liée, dont on suppose que la loi de probabilité n'est pas complètement spécifiée. On se restreint au cas où la forme de la loi est connue à un paramètre près que l'on cherche à estimer.

**Exemple fondamental.** On s'intéresse lors d'un sondage aux intentions de votes de la population. Il y a deux choix lors du vote : une personne $A$ (codée par 1) et une personne $B$ (codée par 0). Si l'on note $X$ la variable aléatoire égale à l'intention de vote d'une personne prise au hasard (uniformément) dans la population, $X$ suit une loi de Bernoulli de paramètre $\theta$ **inconnu**. C'est le paramètre $\theta$ que l'on cherche à estimer via le sondage. Le seul moyen pour connaître la valeur exacte de $\theta$ est de poser la question à toute la population, mais c'est impossible. On cherche donc une valeur approchée de $\theta$ en posant la question à un nombre raisonnable $n$ de personnes. Il reste une question importante ensuite : comment crée-t-on une approximation de $\theta$ à partir des $n$ réponses obtenues ? Répondre à cette question revient à construire un estimateur.

## 2.1 Notion de $n$-échantillon

**Definition 1.** Soit $X$ une variable aléatoire. On appelle *$n$-échantillon* de la variable aléatoire $X$ tout $n$-uplet $(X_1, \dots, X_n)$ de variables aléatoires telles que

1. les variables aléatoires $X_i$ sont mutuellement indépendantes
2. les variables aléatoires $X_i$ suivent toutes la même loi que $X$

*Exemple* 1. Si on lance 10 fois une pièce de monnaie on a naturellement un 10-échantillon de la variable aléatoire $X$ qui donne le résultat d'un lancer (par exemple : 1 si la pièce tombe sur `Pile` et 0 sinon). Il suffit de noter $X_k$ le résultat du $k^{\text{e}}$ lancer et alors $(X_1, \dots, X_{10})$ est un 10-échantillon de $X$.

**Definition 2.** Soit $X$ une variable aléatoire. Soit $(X_1, \dots, X_n)$ un $n$-échantillon de $X$. Pour toute issue $\omega \in \Omega$, on dit que $(X_1(\omega), \dots, X_n(\omega))$ est une *réalisation* (ou *observation*) du $n$-échantillon. On note souvent $(x_1, \dots, x_n)$ une telle réalisation.

## 2.2 Estimateur et estimation

**Definition 3.** Soit $X$ une variable aléatoire dont la loi dépend d'un paramètre $\theta$. Soit $(X_1, \dots, X_n)$ un $n$-échantillon de la variable aléatoire $X$.

1. On appelle *estimateur de $\theta$* toute v.a.r. $T_n$ qui s'exprime en fonction des v.a.r. $(X_1, \dots, X_n)$ et dont l'expression ne fait pas apparaître le paramètre $\theta$.
   Autrement dit, $T_n$ est un estimateur de $\theta$ si il existe une fonction $\varphi$ à $n$ variables telle que :
   $$T_n = \varphi(X_1, \dots, X_n)$$
2. Si $\varphi$ permet de définir un estimateur de $\theta$, alors on appelle *estimation de $\theta$* tout réel de la forme :
   $$\varphi(x_1, \dots, x_n)$$
   où $(x_1, \dots, x_n) \in X_1(\Omega) \times \dots \times X_n(\Omega)$ est une réalisation du $n$-échantillon $(X_1, \dots, X_n)$.
3. Lorsque l'estimateur $T_n$ possède une espérance (resp. une variance), on la note $\mathbb{E}_\theta(T_n)$ (resp. $\mathbb{V}_\theta(T_n)$).
   *(lire espérance / variance sous $\theta$)*

*Remarque* 1. Dans certains cas, on ne cherche pas à estimer le paramètre $\theta$, mais plutôt son image $g(\theta)$ par une certaine fonction $g$. Par exemple, si $X$ suit une loi exponentielle de paramètre inconnu $\lambda$, alors $\mathbb{E}(X) = \dfrac{1}{\lambda}$. Dans ce cas, on peut vouloir trouver une estimation de $g(\lambda) = \dfrac{1}{\lambda}$ plutôt que $\lambda$.

Le fait de noter $\mathbb{E}_\theta(T_n)$ l'espérance de $T_n$ rappelle que celle-ci dépend a priori du paramètre $\theta$. Il en va évidemment de même de la loi de probabilité de la variable $T_n$ : un même événement (par exemple le succès dans une épreuve de Bernoulli) ayant une probabilité différente selon la valeur de $\theta$ ; en toute rigueur, il convient donc de l'indiquer en notant $\mathbb{P}_\theta$ la probabilité en jeu, car elle dépend à son tour de la valeur du paramètre estimé.

## 2.3 L'estimateur moyenne empirique

**Definition 4.** Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$. Soit $n \in \mathbb{N}^*$. Soit $(X_1, \dots, X_n)$ un $n$-échantillon de la v.a.r. $X$.

1. Alors la v.a.r. $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$ est un estimateur de $\theta$.
2. Cet estimateur est appelée *moyenne empirique* de l'échantillon.

La moyenne empirique est un estimateur naturel de $\mathbb{E}_\theta(X)$ du fait de la loi faible des grands nombres.

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

## 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 $\theta$ et que l'on cherche à estimer $\theta$. 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 $\varphi$ est choisie de telle sorte que $\theta^\star = \varphi(x_1, \dots, x_n)$ soit la valeur rendant maximale la probabilité de l'évènement

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

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 $\Theta$ 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 2 :** *(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 $\theta^*$ que l'on explicitera en fonction de $x_1, \dots, x_n$.
3. Montrer que $L_n$ admet également un maximum en $\theta^*$.
4. Expliciter l'estimateur du maximum de vraisemblance. Quel estimateur reconnaît-on ?

**Exercice 3 :** *(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* 2. 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 4 :** *(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 :
$$\hat{\theta}_n = \frac{n}{X_1 + \cdots + X_n}$$

### 2.4.2 Le cas à densité

On peut aussi définir la vraisemblance d'une loi à densité. En notant $f_\theta$ la densité d'une variable aléatoire $X$ de loi de paramètre inconnue $\theta$, 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 5 :** On considère un $n$-échantillon de la loi $\mathcal{U}([0,\theta])$ et on cherche à estimer $\theta$. Soit $(x_1, \dots, x_n) \in (\mathbb{R}_+^*)^n$. On note $f_\theta$ 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 6 :** `(extrait de EDHEC 2012)`

Soit $Y \hookrightarrow \mathcal{E}(\lambda)$ de densité $f_Y$. On suppose, dans la suite, que le paramètre $\lambda$ 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 $Y_1, \ldots, Y_n$, 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 $x_1, \ldots, x_n$ 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 $\lambda$, $x_1, \ldots, x_n$.

   (b) On considère la fonction $\varphi$, définie pour tout réel $\lambda$ de $]0,+\infty[$ par : $\varphi(\lambda) = n\ln(\lambda) - \lambda \sum_{k=1}^{n} x_k$.

   Montrer que la fonction $\varphi$ admet un maximum, atteint en un seul réel que l'on notera $z$ et que l'on exprimera en fonction de $x_1, \ldots, x_n$.
   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 $\lambda$.

   (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 $\lambda$, fonction simple de $Z_n$, qui soit un estimateur sans biais de $\lambda$ (*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 $\theta$. 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 $\theta$ 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 $\theta$
- 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 $\alpha$ 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* 3. 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 $\mathbb{P}_\theta\Big([u_n \leq g(\theta) \leq v_n]\Big) = 0$.

**A retenir.**

| 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 $([U_n, V_n])_{n\in\mathbb{N}^*}$ vérifiant : il existe une suite de réels $(\alpha_n)_{n\in\mathbb{N}^*}$ à valeurs dans $[0,1]$, de limite $\alpha$, 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* 4. 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.

**A retenir.**

| 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 $p$ est inconnu. Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite de variables aléatoires indépendantes de même loi que $X$. On note, pour tout $n \in \mathbb{N}^*$, $\overline{X}_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$.

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 $\varepsilon$ pour que ce soit le cas. La première difficulté vient de la dépendance en $p$ (on ne souhaite pas que $\varepsilon$ 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* 5. Nous pouvons observer sur cet intervalle de confiance deux résultats intuitifs :

- plus le niveau de risque $\alpha$ 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* 2. 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 $\varepsilon$ (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 $\varepsilon$ (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 $\varepsilon$ faible).

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

| $\varepsilon$ \ $1-\alpha$ | 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

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

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

| $\varepsilon$ \ $1-\alpha$ | 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 $t_\alpha$ 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 $\left[\overline{x_n} - \varepsilon, \overline{x_n} + \varepsilon\right]$.
  - 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 $n = 1537$ personnes, il assure avec une probabilité de 95% qu'un candidat est évalué à 19% plus ou moins 2,5%. 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 7 :** 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 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 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 $\alpha$.*

# 4 Compléments hors-programme sur l'estimation ponctuelle et la comparaison d'estimateurs

On développe dans cette partie quelques notions qui permettent de quantifier la *qualité* d'un estimateur. Une fois cette quantification faite, on peut comparer deux estimateurs et choisir celui qui est le plus pertinent.

## 4.1 Biais d'un estimateur

**Definition 8.** Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$. Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$.

1. Si $T_n$ admet une espérance, on appelle *biais de l'estimateur* $T_n$ le réel :
$$b_\theta(T_n) = \mathbb{E}_\theta(T_n) - \theta$$
2. Si le paramètre estimé est $g(\theta)$, l'expression du biais devient :
$$b_\theta(T_n) = \mathbb{E}_\theta(T_n) - g(\theta)$$
3. On dit que $T_n$ est un *estimateur sans biais de* $\theta$ lorsque $b_\theta(T_n) = 0$ (*i.e.* $E_\theta(T_n) = \theta$). Dans le cas contraire, on parlera d'estimateur biaisé.

*Exemple* 4.

1. Si $X \hookrightarrow \mathcal{B}(p)$, alors l'estimateur moyenne empirique $\overline{X}_n$ est un estimateur sans biais de $p$.
2. Si $Y \hookrightarrow \mathcal{P}(\lambda)$. Alors :
$$\mathbb{E}(\overline{Y}_n) = \mathbb{E}\left(\frac{1}{n}\sum_{k=1}^{n} Y_k\right) = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(Y_k) = \frac{1}{n}\sum_{k=1}^{n}\lambda = \frac{1}{n}\,n\lambda = \lambda = \mathbb{E}(Y)$$
Ainsi, $\overline{Y}_n$ est un estimateur sans biais de $\lambda$.
3. Soit $X$ une v.a.r. qui suit la loi $\mathcal{E}(\lambda)$ où $\lambda$ est un paramètre inconnu. Soit $(X_1, \dots, X_n)$ un $n$-échantillon de la v.a.r. $X$. Les $U_n = \overline{X}_n$ et $V_n = X_1$ sont des estimateurs sans biais de $\dfrac{1}{\lambda}$.

*Remarque* 6. Le biais mesure l'écart *moyen* entre les valeurs prises par l'estimateur et le paramètre $\theta$ à estimer. Si l'estimateur est sans biais, les valeurs de l'estimateur sont *en moyenne* très proches de $\theta$. Intuitivement, un estimateur sans biais semble de meilleure qualité qu'un estimateur biaisé ; ce n'est cependant pas toujours le cas car des valeurs très éloignées de $\theta$ peuvent donner $\theta$ en moyenne.

**Exercice 9 :** Soit $(X_1, \dots, X_n)$ un $n$-échantillon d'une variable $X \hookrightarrow \mathcal{U}([0, \theta])$ où $\theta > 0$.

1. (a) Montrer que $\overline{X}_n$ est un estimateur biaisé de $\theta$ et préciser $b_\theta(\overline{X}_n)$.
   (b) Proposer alors un estimateur $V_n$ de $\theta$ sans biais, obtenu comme transformation simple de $\overline{X}_n$.
2. On considère maintenant l'estimateur $M_n = \max(X_1, \dots, X_n)$.
   (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 $M_n$.

**Exercice 10 :** *(Estimation de la variance)*

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

1. On suppose que $m$ est connu. Montrer que $T_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$ est un estimateur non biaisé de $\sigma^2$.
2. On suppose que $m$ n'est pas connu. On note $\overline{X}_n$ la moyenne empirique de l'échantillon et
$$U_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

(a) Montrer que, pour tout $i \in [\![1, n]\!]$, $\mathbb{E}((X_i - \overline{X}_n)^2) = \mathbb{V}(X_i - \overline{X}_n)$.

(b) Montrer que, pour tout $i \in [\![1, n]\!]$,

$$\mathbb{V}(X_i - \overline{X}_n) = \left(1 - \frac{1}{n}\right)^2 \mathbb{V}(X_i) + \frac{1}{n^2} \sum_{\substack{1 \leqslant k \leqslant n \\ k \neq i}} \mathbb{V}(X_k).$$

(c) En déduire que

$$\mathbb{V}(X_i - \overline{X}_n) = \frac{n-1}{n}\sigma^2.$$

(d) Montrer alors que $U_n$ est un estimateur biaisé de $\sigma^2$. En déduire un estimateur sans biais de $\sigma^2$.

## 4.2 Estimateur asymptotiquement sans biais

**Definition 9.** Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$. Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $\theta$ (resp. $g(\theta)$).

On dit que la suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ est *asymptotiquement sans biais* lorsque :

$$\lim_{n\to+\infty} b_\theta(T_n) = 0 \quad \text{ou encore} \quad \lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = \theta$$

$$(\text{resp. } \lim_{n\to+\infty} \mathbb{E}_\theta(T_n) = g(\theta))$$

On dit aussi, par abus de langage, que l'estimateur $T_n$ est *asymptotiquement sans biais*.

## 4.3 Risque quadratique

**Definition 10.** Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$. Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$ (resp. $g(\theta)$). Si $T_n$ admet un moment d'ordre 2, alors on appelle *risque quadratique de l'estimateur* $T_n$ le réel :

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

*Remarque* 7.

1. Le risque quadratique mesure la moyenne des carrés des écarts au paramètre $\theta$ (on remarquera la similitude avec la définition de la variance). Comme un carré est toujours positif, les écarts positifs ou négatifs à $\theta$ ne se compensent pas mais se cumulent (contrairement au biais). On a donc bien ici une façon de mesurer si $T_n$ est un « bon » estimateur de $\theta$.
2. $r_\theta(T_n)$ est toujours positif ou nul. Dans le cas où $r_\theta(T_n) = 0$, l'estimateur $T_n$ est presque sûrement égal à $\theta$, ce qui en fait l'estimateur parfait.
3. Entre deux estimateurs de $\theta$, on choisira celui dont le risque quadratique est le plus faible.

**Théoreme 2** (Décomposition biais - variance)**.** *Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$ . Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$ (resp. $g(\theta)$). Si $T_n$ admet un moment d'ordre 2. Alors on a*

$$r_\theta(T_n) = \mathbb{V}_\theta(T_n) + (b_\theta(T_n))^2$$

*Démonstration.* Tout d'abord, comme la v.a.r. $T_n$ admet un moment d'ordre 2, alors elle admet une variance, un risque quadratique et un biais. Ensuite :

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

□

**Corollaire 3.** *Si $T_n$ est un estimateur sans biais, alors son risque quadratique est égal à sa variance : $r_\theta(T_n) = \mathbb{V}_\theta(T_n)$.*

**Exercice 11 :** Soit $(X_1, \dots, X_n)$ un $n$-échantillon de $X \hookrightarrow \mathcal{B}(\theta)$.

1. Calculer le biais et le risque quadratique de l'estimateur $T_n = X_1$.
2. Calculer le biais et le risque quadratique de l'estimateur $\overline{X}_n$.
3. Quel estimateur de $\theta$ choisir ?

**Exercice 12 :** Soit $(X_1, \dots, X_n)$ un $n$-échantillon d'une variable $X \hookrightarrow \mathcal{U}([0, \theta])$ où $\theta > 0$. On a deux estimateurs sans biais de $\theta$ :

$$V_n = \frac{2}{n}(X_1 + \cdots + X_n), \qquad Z_n = \frac{n+1}{n}\max(X_1, \dots, X_n).$$

1. Établir que
$$r_\theta(V_n) = \mathbb{V}_\theta(V_n) = \frac{\theta^2}{3n}.$$
2. Montrer que
$$\mathbb{E}_\theta(Z_n^2) = \frac{(n+1)^2}{n^2}\int_0^\theta \frac{nt^{n+1}}{\theta^n}\mathrm{d}t.$$
En déduire que
$$r_\theta(Z_n) = \mathbb{V}_\theta(Z_n) = \frac{\theta^2}{n(n+2)}$$
3. Quel estimateur aura-t-on tendance à préférer en pratique ?

## 4.4 Estimateur convergent

**Definition 11.** Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$. Soit $(T_n)_{n\in\mathbb{N}^*}$ une suite d'estimateurs de $\theta$ (resp. $g(\theta)$). On dit que la suite d'estimateurs $(T_n)_{n\in\mathbb{N}^*}$ de $\theta$ (resp. $g(\theta)$) est *convergente* si :

$$\forall \varepsilon > 0, \ \lim_{n\to+\infty} \mathbb{P}([|T_n - \theta| > \varepsilon]) = 0$$

$$\text{resp. } \forall \varepsilon > 0, \ \lim_{n\to+\infty} \mathbb{P}([|T_n - g(\theta)| > \varepsilon]) = 0$$

On dit aussi, par abus de langage, que l'estimateur $T_n$ est *convergent*.

*Remarque* 8. Un estimateur convergent s'écarte donc du paramètre à estimer avec très faible probabilité si la taille de l'échantillon est assez grande.

*Remarque* 9. La loi faible des grands nombres implique notamment que la moyenne empirique $\overline{X_n}$ est un estimateur convergent de l'espérance.

**Théoreme 4.** *Soit $X$ une v.a.r. dont la loi dépend d'un paramètre $\theta$. Soit $n \in \mathbb{N}^*$ et soit $T_n$ un estimateur de $\theta$ (resp. $g(\theta)$). On suppose que $T_n$ admet un moment d'ordre 2.*

$$\boxed{\lim_{n\to+\infty} r_\theta(T_n) = 0 \implies T_n \text{ est un estimateur convergent}}$$

*Démonstration.* Soit $\varepsilon > 0$. La v.a.r. $(T_n - \theta)^2$ :

1. admet une espérance,
2. est positive presque sûrement.

Ainsi, par inégalité de Markov (avec $a = \varepsilon^2$) :

$$\mathbb{P}_\theta\left(\left[(T_n-\theta)^2 > \varepsilon^2\right]\right) \leq \frac{\mathbb{E}_\theta((T_n-\theta)^2)}{\varepsilon^2}$$

Or, par stricte croissance de $x \mapsto x^2$ sur $[0,+\infty[$ :

$$\mathbb{P}_\theta([|T_n-\theta| > \varepsilon]) = \mathbb{P}_\theta\left(\left[(T_n-\theta)^2 > \varepsilon^2\right]\right)$$

Donc :

$$0 \leq \mathbb{P}_\theta([|T_n-\theta| > \varepsilon]) \leq \frac{\mathbb{E}_\theta((T_n-\theta)^2)}{\varepsilon^2} = \frac{r_\theta(T_n)}{\varepsilon^2}$$

Or : $\lim\limits_{n\to+\infty} \dfrac{r_\theta(T_n)}{\varepsilon^2} = 0$. Ainsi, par théorème d'encadrement : $\lim\limits_{n\to+\infty} \mathbb{P}_\theta([|T_n-\theta| > \varepsilon]) = 0$. □

# 5 Exercices supplémentaires

## 5.1 Maximum de vraisemblance

**Exercice 13 :** 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 $\theta$ (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 $\theta$, la loi de $X$. On note $p = \mathbb{P}([X = 1])$. Exprimer alors $\theta$ en fonction de $p$.
   (c) Quelle est, en fonction de $\theta$, la probabilité qu'une personne ayant choisi la réponse 1 l'ait fait car elle connaissait réellement la réponse ?
2. Afin d'estimer $\theta$, 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 $\theta$.
   (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 $\theta$. 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 $\theta$ 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.

## 5.2 Intervalles de confiance

**Exercice 14 :** Soit $(X_1, \dots, X_n)$ un $n$-échantillon d'une variable $X \hookrightarrow \mathcal{U}([0, \theta])$ où $\theta > 0$. On a l'estimateur sans biais de $\theta$ :

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

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 $\alpha$ pour $\theta$.

**Exercice 15 :** Dans une population d'un pays totalement fictif, des électeurs doivent choisir le futur président parmi 2 candidats, Jean-Michel Peste et Jean-Pierre Choléra. On note $p$ la proportion d'électeurs désirant voter pour M. Peste. On choisit un échantillon $(X_1, \dots, X_n)$ où l'on a noté $X_i = 1$ si la personne a voté pour M. Peste, et 0 sinon. Parmi ces personnes, 55% déclarent vouloir voter pour M. Peste.

1. Peut-on déclarer au risque $\alpha = 5\%$ que M. Peste sera élu président si $n = 100$ ? On utilisera le théorème central limite.
2. Même question avec $n = 10000$.

**Exercice 16 :** 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 17 :** 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 $\alpha$ 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 $t_\alpha$ le réel défini par $\Phi(t_\alpha) = 1 - \frac{\alpha}{2}$, où $\Phi$ 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 18 :** Soit $X$ une v.a.r. suivant la loi uniforme sur $[0,\theta]$, où $\theta > 0$ est un paramètre inconnu. Soit $(X_1, X_2, \ldots, X_n)$ un $n$-échantillon de la v.a.r. $X$. On considère les estimateurs suivants :

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

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

1. $T_n$ peut-il être un estimateur de $\theta$ ?
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 réalisation 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 $V_n = 1,01\ U_n$ ?

**Exercice 19 :** 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 $(Y_1, Y_2, \ldots, Y_m)$ 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 $\varepsilon$ un réel strictement positif quelconque.*

1. (a) Montrer que $\overline{Y_m}$ est un estimateur sans biais de $p$ ; déterminer son risque quadratique.

   (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 $\theta$ 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 20 :** 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.

**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.

1. 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.
2. 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 $p$.
3. On pose $X_n = \dfrac{m}{n} R_n$. Montrer que $X_n$ est un estimateur sans biais convergent de $N$.
4. (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 $\Phi(2) \simeq 0,975$.
5. 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 $R_n$.
6. En déduire que si $x \in ]0,1[$, alors la série $\sum_i \binom{n+i}{n} x^i$ est convergente, et calculer sa somme.

## 5.3 Estimation ponctuelle, comparaison d'estimateurs

**Exercice 21 :** La durée de vie d'une lampe est une v.a.r. qui suit une loi exponentielle de paramètre $\frac{1}{m}$ inconnu. On cherche à estimer la durée de vie moyenne $m$ de la lampe. On prélève un échantillon de $n$ lampes et on note $X_1, \dots, X_n$ leurs durées de vie. On pose $\overline{X}_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$.

1. (a) Montrer que $\overline{X}_n$ est un estimateur sans biais de $m$.
   (b) Calculer son risque quadratique.
2. On pose $Y_n = \min(X_1, \dots, X_n)$.
   (a) Déterminer la loi de $Y_n$.
   (b) En déduire que $Z_n = nY_n$ est un estimateur sans biais de $m$.
   (c) Calculer son risque quadratique.
3. Comparer les deux estimateurs.

**Exercice 22 :** La sécurité routière fait une enquête sur le nombre d'accidents survenus par semaine sur un tronçon d'autoroute. Soit $X$ la v.a.r. égale au nombre d'accidents par semaine. On suppose que $X$ suit une loi de Poisson de paramètre $\theta$ inconnu ($\theta \in ]0, +\infty[$). On se propose d'évaluer le paramètre $e^{-\theta} = \mathbb{P}([X = 0])$. On note $X_1$, $X_2$, …, $X_n$ les résultats des observations faites pendant $n$ semaines. On suppose $X_1, \dots, X_n$ indépendantes et de même loi que $X$.

1. Pour tout $i \in [\![1, n]\!]$, on définit $Y_i$ par : $Y_i = 1$ si $X_i = 0$, et $Y_i = 0$ sinon. On note aussi : $\overline{Y_n} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$.
   (a) Pour tout $i \in [\![1, n]\!]$, donner la loi de $Y_i$.
   (b) Montrer que $\overline{Y_n}$ est un estimateur sans biais de $e^{-\theta}$.
   (c) Calculer le risque quadratique de $\overline{Y_n}$.
   (d) Montrer que $\overline{Y_n}$ est un estimateur convergent de $e^{-\theta}$.
   (e) Expliquer pourquoi $\overline{Y_n}$ est un estimateur « naturel » de $e^{-\theta}$.

   Cet estimateur ne tient pas compte du fait que $X$ suit une loi de Poisson. On peut donc espérer trouver un meilleur estimateur sans biais convergent de $e^{-\theta}$.
2. On pose $S_n = \sum_{i=1}^{n} X_i$.
   (a) Quelle est la loi de $S_n$ ?
   (b) Calculer l'espérance de $e^{-\frac{S_n}{n}}$ à l'aide du théorème de transfert.
   (c) Montrer que $e^{-\frac{S_n}{n}}$ est un estimateur biaisé de $e^{-\theta}$.
   (d) Montrer que $e^{-\frac{S_n}{n}}$ est asymptotiquement sans biais, c'est-à-dire que $\lim\limits_{n\to+\infty} \mathbb{E}\left(e^{-\frac{S_n}{n}}\right) = e^{-\theta}$.
3. Pour tout entier naturel $j$, on définit la probabilité conditionnelle :

   $$\varphi(j) = \mathbb{P}_{[S_n = j]}([X_1 = 0])$$

   Montrer que, pour tout $j \in \mathbb{N}$, $\varphi(j) = \left(1 - \dfrac{1}{n}\right)^j$. On a donc $\varphi(j)$ indépendant du paramètre $\theta$ inconnu.
4. (a) Montrer que $\varphi(S_n)$ est un estimateur sans biais de $e^{-\theta}$.
   (b) Calculer le risque quadratique de $\varphi(S_n)$.
   (c) Montrer que $\varphi(S_n)$ est un estimateur convergent de $e^{-\theta}$.
5. (a) En utilisant le théorème des accroissements finis, démontrer que : $1 \leq \dfrac{\exp(\theta) - 1}{\theta} \leq \exp(\theta)$.
   (b) Soit $h$ la fonction définie sur $[0, 1]$ par $h(t) = t\exp(\theta) + (1 - t) - \exp(t\theta)$. Étudier les variations de $h$.
   (c) En déduire que : $\exp\left(\dfrac{\theta}{n}\right) \leq \dfrac{\exp(\theta)}{n} + \dfrac{n-1}{n}$.
   (d) Quel est le meilleur estimateur de $e^{-\theta}$ entre $\varphi(S_n)$ et $\overline{Y_n}$ ?

**Exercice 23 :** Soient $X_1, \dots, X_n$, $Y_1, \dots, Y_m$, $n+m$ variables aléatoires indépendantes suivant une loi de Bernoulli de paramètre inconnu $p$. On se propose d'estimer $p$. On suppose dans la suite que $n > m$. On considère les estimateurs de $p$ suivants :

$$M_1 = \frac{1}{n}\sum_{k=1}^{n} X_k,\ M_2 = \frac{1}{m}\sum_{k=1}^{m} Y_k \ \text{ et } \ N = \frac{M_1 + M_2}{2}$$

1. (a) Déterminer le biais des estimateurs $M_1$, $M_2$ et $N$.
   (b) Démontrer que ces 3 estimateurs sont convergents.
   (c) Quel est le meilleur des trois estimateurs ? On discutera suivant les valeurs de $n$ et $m$.
2. On considère des estimateurs de $p$ de la forme $aM_1 + bM_2$ avec $(a,b) \in \mathbb{R}^2$.
   (a) Parmi ces estimateurs, lequel est le meilleur estimateur sans biais ?
   (b) Quel est son risque quadratique ?

**Exercice 24 :** On considère un réel $r > 0$ et la fonction $f$ suivante :

$$f : t \mapsto \begin{cases} \dfrac{2t}{r^2} & \text{si } t \in [0,r] \\ 0 & \text{si } t \notin [0,r] \end{cases}$$

1. (a) Étudier la continuité de $f$.
   (b) Montrer que $f$ est une densité de probabilité.

   On note dans toute la suite $X$ une v.a.r. réelle de densité $f$. $F_X$ désigne sa fonction de répartition.
2. (a) Déterminer la valeur $F_X(x)$ lorsque $x < 0$, puis lorsque $x > r$.
   (b) Montrer que pour tout réel $x$ de $[0,r]$, $F_X(x) = \dfrac{x^2}{r^2}$.
3. (a) Montrer que $X$ admet une espérance et que $\mathbb{E}(X) = \dfrac{2r}{3}$.
   (b) Montrer que $X$ admet une variance et que $\mathbb{V}(X) = \dfrac{r^2}{18}$.

   Dans toute la suite $n$ désigne un entier naturel non nul et $(X_1, X_2, \dots, X_n)$ un $n$-échantillon de la v.a.r. $X$. On cherche alors à estimer le réel $r$.
4. On note $T_n = \dfrac{3}{2n}\sum_{k=1}^{n} X_k$ et on cherche à estimer $r$ avec $T_n$.
   (a) Montrer que $T_n$ est un estimateur sans biais de $r$.
   (b) Calculer le risque quadratique de $T_n$ (noté $r(T_n)$).
5. On note $M_n$ la v.a.r. prenant pour valeur le maximum des valeurs prises par les variables $X_1, X_2, \dots, X_n$, de sorte que, pour tout réel $x$ :

$$[M_n \leq x] = [X_1 \leq x] \cap [X_2 \leq x] \cap \dots \cap [X_n \leq x]$$

   (a) Montrer que : $\forall x \in \mathbb{R}$, $\mathbb{P}([M_n \leq x]) = (F_X(x))^n$. En déduire la fonction de répartition de $M_n$. Puis montrer que $M_n$ est une v.a.r. à densité.
   (b) Montrer qu'une densité possible de $M_n$ est la fonction $g_n$ définie par :

$$g_n : t \mapsto \begin{cases} 2n\,\dfrac{t^{2n-1}}{r^{2n}} & \text{si } t \in [0,r] \\ 0 & \text{si } t \notin [0,r] \end{cases}$$

   (c) Montrer que $M_n$ admet une espérance et une variance, et que :

$$\mathbb{E}(M_n) = \frac{2n}{2n+1}\,r \quad \text{et} \quad \mathbb{V}(M_n) = \frac{n}{(n+1)(2n+1)^2}r^2$$

   (d) On cherche à estimer $r$ avec $M_n$. Calculer le biais de $M_n$, noté $b(M_n)$, et son risque quadratique $r(M_n)$.
6. (a) Déterminer un équivalent simple lorsque $n \to +\infty$ de $b(M_n)$ et $r(M_n)$.

(b) Quels sont les avantages et les inconvénients réciproques des estimateurs $T_n$ et $M_n$ ?

7. Déduire de $M_n$ un estimateur $U_n$ sans biais de $r$. Entre $T_n$ et $U_n$, quel estimateur de $r$ choisissez-vous ?
8. On dit qu'un estimateur $Z_n$ de $r$ est convergent lorsque : $\forall \varepsilon > 0, \lim\limits_{n\to+\infty} \mathbb{P}(|Z_n - r| > \varepsilon) = 0$. Dans cette question, on montre que les estimateurs $T_n, M_n$ et $U_n$ sont des estimateurs convergents de $r$.

(a) Montrer que $T_n$ est un estimateur convergent de $r$.

(b) i. Montrer que, pour tout $\varepsilon \in ]0, r[$ : $[|M_n - r| > \varepsilon] = [M_n - r < -\varepsilon]$.

ii. En déduire que, pour tout $\varepsilon \in ]0, r[$ :

$$\lim_{n\to+\infty} \mathbb{P}([|M_n - r| > \varepsilon]) = 0$$

Puis montrer que $M_n$ est un estimateur convergent.

(c) i. Montrer que, pour tout $\varepsilon \in ]0, r[$ :

$$[|U_n - r| > \varepsilon] = \left[M_n > \frac{2n(r+\varepsilon)}{2n+1}\right] \bigcup \left[M_n < \frac{2n(r-\varepsilon)}{2n+1}\right]$$

ii. En déduire que, pour tout $\varepsilon \in ]0, r[$ :

$$\mathbb{P}([|U_n - r| > \varepsilon]) = \mathbb{P}\left(\left[M_n > \frac{2n(r+\varepsilon)}{2n+1}\right]\right) + \left(\frac{2n}{2n+1}\right)^{2n} \left(1 - \frac{\varepsilon}{r}\right)^{2n}$$

iii. En déduire que, pour tout $\varepsilon \in ]0, r[$ : $\lim\limits_{n\to+\infty} \mathbb{P}([|U_n - r| > \varepsilon]) = 0$. Puis montrer que $U_n$ est un estimateur convergent.

(d) Montrer que : $M_n \xrightarrow[n\to+\infty]{\mathscr{L}} r$.

# 6 Appendice : Table de la loi normale centrée réduite $\mathcal{N}(0,1)$

| $x$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.5000 | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | 0.9495 | 0.9505 | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | 0.9750 | 0.9756 | 0.9761 | 0.9767 |
| 2.0 | 0.9772 | 0.9778 | 0.9783 | 0.9788 | 0.9793 | 0.9798 | 0.9803 | 0.9808 | 0.9812 | 0.9817 |
| 2.1 | 0.9821 | 0.9826 | 0.9830 | 0.9834 | 0.9838 | 0.9842 | 0.9846 | 0.9850 | 0.9854 | 0.9857 |
| 2.2 | 0.9861 | 0.9864 | 0.9868 | 0.9871 | 0.9875 | 0.9878 | 0.9881 | 0.9884 | 0.9887 | 0.9890 |
| 2.3 | 0.9893 | 0.9896 | 0.9898 | 0.9901 | 0.9904 | 0.9906 | 0.9909 | 0.9911 | 0.9913 | 0.9916 |
| 2.4 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 | 0.9929 | 0.9931 | 0.9932 | 0.9934 | 0.9936 |
| 2.5 | 0.9938 | 0.9940 | 0.9941 | 0.9943 | 0.9945 | 0.9946 | 0.9948 | 0.9949 | 0.9951 | 0.9952 |
| 2.6 | 0.9953 | 0.9955 | 0.9956 | 0.9957 | 0.9959 | 0.9960 | 0.9961 | 0.9962 | 0.9963 | 0.9964 |
| 2.7 | 0.9965 | 0.9966 | 0.9967 | 0.9968 | 0.9969 | 0.9970 | 0.9971 | 0.9972 | 0.9973 | 0.9974 |
| 2.8 | 0.9974 | 0.9975 | 0.9976 | 0.9977 | 0.9977 | 0.9978 | 0.9979 | 0.9979 | 0.9980 | 0.9981 |
| 2.9 | 0.9981 | 0.9982 | 0.9982 | 0.9983 | 0.9984 | 0.9984 | 0.9985 | 0.9985 | 0.9986 | 0.9986 |
| 3.0 | 0.9987 | 0.9987 | 0.9987 | 0.9988 | 0.9988 | 0.9989 | 0.9989 | 0.9989 | 0.9990 | 0.9990 |
| 3.1 | 0.9990 | 0.9991 | 0.9991 | 0.9991 | 0.9992 | 0.9992 | 0.9992 | 0.9992 | 0.9993 | 0.9993 |
| 3.2 | 0.9993 | 0.9993 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9994 | 0.9995 | 0.9995 | 0.9995 |
| 3.3 | 0.9995 | 0.9995 | 0.9995 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9996 | 0.9997 |
| 3.4 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9997 | 0.9998 |
| 3.5 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.9998 |
| 3.6 | 0.9998 | 0.9998 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.7 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.8 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 | 0.9999 |
| 3.9 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |