Feuille d'exercices n°15 : Estimation

Estimation ponctuelle

Exercice 1. (★)
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, X_2, $\dots$, X_n leurs durées de vie. On pose $\overline{X}_n = \frac{1}{n}\sum\limits_{k=1}^{n} X_k$.

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 2. (☆)
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 de la limite centrée.

Exercice 3. (★)
Soit X une v.a.r. suivant la loi uniforme sur $[0, \theta]$, où $\theta > 0$ est un paramètre inconnu. On considère un n-échantillon $(X_1, X_2, \dots, X_n)$ de la v.a.r. X.
On considère les estimateurs suivants :

$$T_n = 2\overline{X_n}, \quad T_n' = \max(X_1, X_2, ..., X_n) \quad \text{et} \quad T_n'' = \frac{n+1}{n}\, T_n'$$

où $\overline{X_n} = \frac{1}{n}\sum\limits_{k=1}^{n} X_k$ est la moyenne empirique du n-échantillon.

1. *a)* Montrer que T_n est un estimateur sans biais de θ.
b) Déterminer le risque quadratique de T_n.
c) T_n est-il un estimateur convergent de θ ?

2. *a)* Déterminer les biais et risque quadratique de T_n'.
b) Donner un équivalent simple en $+\infty$ de $r_\theta(T_n')$.

3. Mêmes questions pour T_n''.
Quel est le meilleur des trois estimateurs ?

4. Écrire un programme **Scilab** simulant ces trois estimateurs.

Exercice 4. (★)
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 5. (★★)
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,
  × 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 $1 \leqslant i \leqslant n$, 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 : $\forall \varepsilon > 0,\ \mathbb{P}([|T_n - \theta| \geqslant \varepsilon]) \leqslant \dfrac{1}{20\ n\ \varepsilon^2}$.
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 $\widehat{\theta}$ de θ est alors fournie par une réalisation de T_n.

Dans cette question, on cherche à obtenir une estimation $\widehat{p}$ 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 $\widehat{p}$ le point en lequel f atteint ce maximum.

Remarque
La quantité $\widehat{p}$ s'écrit sous la forme :

$$\widehat{p} = h(v_1, \ldots, v_n)$$

Elle est choisie comme estimation du paramètre p.
On définit alors L'estimateur Z_n du **maximum de vraisemblance** par :

$$Z_n = h(V_1, \ldots, V_n)$$

Exercice 6. (★) *(d'après* `ESC 2006`*)*
Dans cet exercice, on considère $r > 0$ 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 $(X_1, X_2, \ldots, X_n)$ un n-échantillon de la v.a.r. X. On cherche alors à estimer le réel r.

4. On note $T_n = \dfrac{3}{2n} \sum\limits_{k=1}^{n} X_k$ 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 $X_1, X_2, \ldots, X_n$, de sorte que, pour tout réel x :

$$[M_n \leqslant x] \ = \ [X_1 \leqslant x] \cap [X_2 \leqslant x] \cap \ldots \cap [X_n \leqslant x]$$

a) Montrer que : $\forall x \in \mathbb{R}, \mathbb{P}([M_n \leqslant x]) = (F_X(x))^n$.
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. Montrer que T_n est un estimateur convergent de r.

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

b) 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.

10. a) Montrer que, pour tout $\varepsilon \in \,]0, r[$:

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

b) En déduire que, pour tout $\varepsilon \in \,]0, r[$:

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

c) 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.

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

Exercice 7. (★)
Soit X une v.a.r. réelle admettant une espérance est m et une variance σ^2.
Soit $(X_1, X_2, \dots, X_n)$ un n-échantillon de la v.a.r. X.

1. Montrer que $T_n = \frac{1}{n} \sum\limits_{k=1}^{n} X_k$ est un estimateur sans biais de m.

2. On pose $V_n = \frac{1}{n} \sum\limits_{k=1}^{n} (X_k - T_n)^2$.

a) Montrer que : $V_n = \frac{1}{n} \left(\sum\limits_{k=1}^{n} X_k^2 \right) - (T_n)^2$.

b) Montrer que : $V_n = \frac{1}{n} \left(\sum\limits_{k=1}^{n} (X_k - m)^2 \right) - (T_n - m)^2$.

c) Montrer que $\mathbb{E}(V_n) = \frac{n-1}{n} \sigma^2$.

d) Construire, à partir de V_n, un estimateur sans biais $\widehat{V_n}$ de σ^2.

e) On dispose d'un échantillon de n observations $(x_1, x_2, \dots, x_n)$ de la v.a.r. X. Donner une méthode pour obtenir une estimation ponctuelle de σ^2 à partir de ces observations.

Exercice 8. (★★) *(adapté de* `ESSEC 2009 - Maths II`*)*
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 θ 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 : $\overline{Y_n} = \frac{1}{n} \sum\limits_{i=1}^{n} Y_i$.

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 $S_n = \sum\limits_{i=1}^{n} X_i$.

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 $j \in \mathbb{N}$, $\varphi(j) = \left(1 - \frac{1}{n}\right)^j$.

On a donc $\varphi(j)$ indépendant du paramètre θ 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 :

$$1 \leqslant \frac{\exp(\theta) - 1}{\theta} \leqslant \exp(\theta)$$

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 : $\exp\left(\frac{\theta}{n}\right) \leqslant \frac{\exp(\theta)}{n} + \frac{n-1}{n}$.

d) Quel est le meilleur estimateur de $e^{-\theta}$ entre $\varphi(S_n)$ et $\overline{Y_n}$?

Exercice 9. (★) (extrait de `ESSEC 2007 - Maths III`)
Dans tout l'exercice, les variables aléatoires sont définies sur un même espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$. On admet le résultat suivant :
si $(X_n)_{n\geqslant 1}$ est une suite de variables aléatoires indépendantes, de même loi exponentielle de paramètre λ (avec $\lambda > 0$), alors pour tout entier naturel n non nul, la v.a.r. $S_n = \sum\limits_{k=1}^{n} X_k$ est une v.a.r. à densité, de densité la fonction f_n définie sur $\mathbb{R}$ par : $f_n : t \mapsto \begin{cases} 0 & \text{si } t < 0 \\ \dfrac{\lambda^n}{(n-1)!}\, \mathrm{e}^{-\lambda t}\, t^{n-1} & \text{si } t \geqslant 0 \end{cases}$

Loi de Pareto

(Vilfredo Pareto (1848-1923), sociologue et économiste italien)
Soient a et b des réels strictement positifs.
Par définition, on dit d'une v.a.r. qu'elle suit la loi de Pareto de paramètres a et b si elle admet pour densité la fonction f définie sur $\mathbb{R}$ par :

$$f(t) = \begin{cases} 0 & \text{si } t < b \\ a\dfrac{b^a}{t^{a+1}} & \text{si } t \geqslant b. \end{cases}$$

Soit alors X une v.a.r. de loi de Pareto de paramètres a et b.
On admet alors les résultats suivants :

(i) f est bien une densité de probabilité.

(ii) X admet une espérance si, et seulement si $a > 1$.
Dans ce cas, $\mathbb{E}(X) = \dfrac{ab}{a-1}$.

(iii) X admet une variance si, et seulement si $a > 2$.
Dans ce cas, $\mathbb{V}(X) = \dfrac{ab^2}{(a-2)(a-1)^2}$.

(iv) La fonction de répartition F de X est définie sur $\mathbb{R}$ par :

$$\forall x \in \mathbb{R},\ F(x) = \begin{cases} 0 & \text{si } x < b \\ 1 - \left(\dfrac{b}{x}\right)^a & \text{si } x \geqslant b \end{cases}$$

(v) La v.a.r. $Y = \ln\left(\frac{X}{b}\right)$ suit la loi exponentielle de paramètre a.

Estimation des paramètres d'une loi de Pareto

- Les instants aléatoires des arrivées de paquets (symboles binaires représentant de l'information de type audio, vidéo, données, ...) dans un canal de communication sont modélisés par une variable aléatoire X suivant une loi de Pareto de paramètres α et β ($\alpha > 0$, $\beta > 0$).
- Soit $(X_n)_{n\geqslant 1}$ une suite de variables aléatoires indépendantes, toutes de même loi que X.

On suppose tout d'abord que le paramètre β fait partie des caractéristiques connues du canal de communication ; on se propose de déterminer un estimateur de α par une méthode dite « du maximum de vraisemblance ». Pour cela, n désignant un entier naturel non nul et $x_1, \dots, x_n$ des réels supérieurs ou égaux à β, on introduit la fonction $\mathscr{L}$, à valeurs dans $\mathbb{R}$ et définie sur $]0, +\infty[$ par :

$$\mathscr{L}(a) = f_a(x_1) \times \cdots \times f_a(x_n) = \prod_{k=1}^{n} f_a(x_k)$$

où f_a est la fonction définie sur $\mathbb{R}$ par :

$$f_a(t) = \begin{cases} 0 & \text{si } t < \beta \\ a\dfrac{\beta^a}{t^{a+1}} & \text{si } t \geqslant \beta \end{cases}$$

1. Exprimer $\mathscr{L}(a)$, puis $\ln(\mathscr{L}(a))$.

2. On considère la fonction φ de $]0, +\infty[$ dans $\mathbb{R}$:

$$a \mapsto n\ln(a) + na\ln(\beta) - (a+1)\sum_{k=1}^{n} \ln(x_k)$$

a) Démontrer que la fonction φ admet un maximum, atteint en un seul réel que l'on notera w.

b) Exprimer w en fonction de x_1, ..., x_n.

c) Que peut-on dire de w pour la fonction $\mathscr{L}$?

3. On pose dorénavant, pour tout n de $\mathbb{N}^*$:

$$W_n = \frac{n}{\sum\limits_{k=1}^{n} \ln\left(\frac{X_k}{\beta}\right)}$$

(La suite $(W_n)_{n\geqslant 1}$ est appelée *estimateur du maximum de vraisemblance*.)

a) Justifier que la v.a.r. $\sum\limits_{k=1}^{n} \ln\left(\frac{X_k}{\beta}\right)$ admet pour densité la fonction f_n définie dans le préambule en prenant $\lambda = \alpha$.

b) à l'aide du théorème de transfert, en déduire que W_n admet pour espérance $\dfrac{n\alpha}{n-1}$ lorsque $n \geqslant 2$, puis proposer un estimateur sans biais de α construit sur W_n.

4. On pose, pour tout n de $\mathbb{N}^*$:

$$W'_n = \frac{n-1}{n} W_n.$$

a) Soit n un entier supérieur ou égal à 3.

En admettant que le moment d'ordre 2 de W'_n est égal à $\dfrac{(n-1)\alpha^2}{n-2}$, calculer la variance de W'_n puis établir, à l'aide de l'inégalité de Bienaymé-Tchebychev, que pour tout réel ε strictement positif, on a l'inégalité :

$$\mathbb{P}\left(\left[W'_n - \varepsilon < \alpha < W'_n + \varepsilon\right]\right) \geqslant 1 - \frac{\alpha^2}{\varepsilon^2(n-2)}$$

b) On suppose dans cette question (et elle seule) que α est strictement compris entre 1 et 2.

Déterminer un entier naturel N tel que, pour tout entier n supérieur ou égal à N, $\left[W'_n - \frac{1}{10}, W'_n + \frac{1}{10}\right]$ soit un intervalle de confiance du paramètre α au niveau de confiance 0,95.

On suppose maintenant que seul le paramètre α est déjà identifié et qu'il vérifie : $\alpha > 2$.

5. Pour tout entier strictement positif n, on pose :

$$Y_n = c_n \sum_{k=1}^{n} X_k,$$

où le réel c_n est choisi de sorte que $(Y_n)_{n\geqslant 1}$ soit un estimateur sans biais de β.

a) Calculer c_n.

b) Quelle est la limite de la variance de Y_n quand n tend vers $+\infty$? Qu'en déduire ?

6. Pour tout entier strictement positif n, on pose : $Z_n = \min(X_1, \dots, X_n)$.

a) Déterminer la fonction de répartition de Z_n, puis reconnaître sa loi et préciser son espérance. Quelle est la limite de cette dernière quand n tend vers $+\infty$?

b) Pour tout entier strictement positif n, on pose : $Z'_n = d_n Z_n$, où le réel d_n est choisi de telle sorte que $(Z'_n)_{n\geqslant 1}$ soit un estimateur sans biais de β.

Quelle est la limite de la variance de Z'_n quand n tend vers $+\infty$?

c) Démontrer que l'estimateur $(Z'_n)_{n\geqslant 1}$ est plus efficace que l'estimateur $(Y_n)_{n\geqslant 1}$, c'est-à-dire, qu'à partir d'un certain rang, la variance de Z'_n est inférieure à celle de Y_n.

Intervalles de confiance

Exercice 10. (★) *(d'après* `ESSEC - Maths II - 2005`*)*
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 $n \geqslant 1$.
Pour tout $i \in [\![1, n]\!]$, soit X_i la v.a.r. de Bernoulli définie par :

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

On suppose que dans les conditions de prélèvement, les variables aléatoires $X_1, \dots, X_n$ sont indépendantes. On pose $S_n = \sum\limits_{k=1}^{n} X_k$.

1. a) Montrer que $F_n = \dfrac{S_n}{n}$ 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 $\left(\sqrt{n}\dfrac{F_n - p}{\sqrt{p(1-p)}}\right)_{n \in \mathbb{N}^*}$?

b) Soit f_n la réalisation de F_n sur l'échantillon considéré.
Soit t_α le réel défini par $\Phi(t_\alpha) = 1 - \dfrac{\alpha}{2}$, 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 11. (★★)
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 estimateurs suivants :

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

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 $[U_n \leqslant \theta \leqslant V_n]$ et $\left[0 \leqslant T_n \leqslant n\left(1 - \dfrac{U_n}{V_n}\right)\right]$.

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 :

```
n = input('Valeur de n?')
theta = 5*rand()
for i=1:n
    disp(grand(1,1,'unf',0,theta))
end
```

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 12. ($\bigstar$) *(extrait de* `HEC 2008 - Maths III`*)*
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 :

× chaque jour n, chaque individu sain peut être contaminé par par n'importe lequel des individus contagieux avec la même probabilité p, ces contaminations éventuelles étant indépendantes les unes des autres ;

× un individu contaminé le jour n devient contagieux le jour $n+1$;

× chaque individu contagieux le jour n redevient sain le jour $n+1$.

Estimations ponctuelles et par intervalle de confiance de p

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 : $\overline{Y_m} = \dfrac{1}{m} \sum\limits_{k=1}^{m} Y_k$.
Dans toute cette partie, on note ε un réel strictement positif quelconque.

1. a) Montrer que $\overline{Y_m}$ est un estimateur 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 θ un réel strictement positif.

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

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 :
$$\mathbb{P}([T \geqslant a]) \leqslant \frac{\mathbb{E}(T)}{a}$$

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 :
$$\mathbb{P}([\overline{Y_m} - p \geqslant \varepsilon]) \leqslant \mathrm{e}^{m(g(\theta) - \theta(p+\varepsilon))}$$

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 : $|g''(x)| \leqslant \dfrac{1}{4}$.

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

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

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

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

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 13. (★)
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 $\varepsilon > 0$, $\lim\limits_{n \to +\infty} \mathbb{P}\left(\left[\left|\dfrac{1}{nm} I_n - \dfrac{1}{N}\right| > \varepsilon\right]\right) = 0$.

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 : $\dfrac{1}{k+1}\dbinom{n}{k} = \dfrac{1}{n+1}\dbinom{n+1}{k+1}$.

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 $k \geqslant 2$, $D_k = R_k - R_{k-1}$.
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 $D_1, D_2, \dots, D_k, \dots$ 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 $1 \leqslant k \leqslant n$.
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 $p \geqslant 0,2$. 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 $\sum\limits_i \binom{n+i}{n} x^i$ est convergente, et calculer sa somme.